18.175 Theory of Probability
Fall 2008

Contents

1 Probability Spaces, Properties of Probability. 1

2 Random variables and their properties. Expectation. 4

3 Kolmogorov's Theorem about consistent distributions. 10

4 Laws of Large Numbers. 12

5 Bernstein Polynomials. Hausdorff and de Finetti theorems. 16

6 0 - 1 Laws. Convergence of random series. 21

7 Stopping times, Wald's identity. Another proof of SLLN. 26

8 Convergence of Laws. Selection Theorem. 29

9 Characteristic Functions. Central Limit Theorem on $\mathbb{R}$. 34

10 Multivariate normal distributions and CLT. 38

11 Lindeberg's CLT. Levy's Equivalence Theorem. Three Series Theorem. 42

12 Levy's Continuity Theorem. Poisson Approximation. Conditional Expectation. 46

13 Martingales. Doob's Decomposition. Uniform Integrability. 51

14 Optional stopping. Inequalities for martingales. 55

15 Convergence of martingales. Fundamental Wald's identity. 59

16 Convergence on metric spaces. Portmanteau Theorem. Lipschitz Functions. 65

17 Metrics for convergence of laws. Empirical measures. 70

18 Convergence and uniform tightness. 74

19 Strassen's Theorem. Relationships between metrics. 76

20 Kantorovich-Rubinstein Theorem. 82

21 Prekopa-Leindler inequality, entropy and concentration. 88

22 Stochastic Processes. Brownian Motion. **96**

23 Donsker Invariance Principle. **100**

24 Empirical process and Kolmogorov's chaining. **103**

25 Markov property of Brownian motion. Reflection principles. **109**

26 Laws of Brownian motion at stopping times. Skorohod's imbedding. **114**

List of Figures

2.1 A random variable defined by quantile transformation. 5
2.2 $\sigma(X)$ generated by X. 5
2.3 Pairwise independent but not independent r.v.s. 7

5.1 Polya urn model. 19

7.1 A sequence of stopping times. 28

8.1 Approximating indicator. 29

14.1 Stopping times of level crossings. 57

25.1 Reflecting the Brownian motion. 111

List of Tables

Section 1

Probability Spaces, Properties of Probability.

A pair $(\Omega, \mathcal{A})$ is a *measurable space* if $\mathcal{A}$ is a σ-algebra of subsets of Ω. A collection A of subsets of Ω is an algebra (ring) if:

1. $\Omega \in A$.
2. $C, B \in A \Longrightarrow C \cap B, C \cup B \in A$.
3. $B \in A \Longrightarrow \Omega \setminus B \in A$.
4. A is a σ-algebra, if in addition, $C_i \in A, \forall i \geq 1 \Longrightarrow \bigcup_{i \geq 1} C_i \in A$.

$(\Omega, \mathcal{A}, \mathbb{P})$ is a *probability space* if $\mathbb{P}$ is a probability measure on $\mathcal{A}$, i.e.

1. $\mathbb{P}(\Omega) = 1$.
2. $\mathbb{P}(A) \geq 0, A \in \mathcal{A}$.
3. $\mathbb{P}$ is countably additive: $A_i \in \mathcal{A}, \forall i \geq 1, A_i \cap A_j = \emptyset \ \forall i \neq j \Longrightarrow \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

An equivalent formulation of Property 3 is:

3′. $\mathbb{P}$ is a finitely additive measure and

$$B_n \supseteq B_{n+1}, \bigcap_{n \geq 1} B_n = B \Longrightarrow \mathbb{P}(B) = \lim_n \mathbb{P}(B_n).$$

Lemma 1 *Properties* 3 *and* 3′ *are equivalent.*

Proof.

$3 \Longrightarrow 3'$: Let $C_n = B_n \setminus B_{n+1}$, then $B_n = B \bigcup \Big(\bigcup_{k \geq n} C_k \Big)$ - all disjoint.

By 3, $\mathbb{P}(B_n) = \mathbb{P}(B) + \sum_{k \geq n} \mathbb{P}(C_k) \to \mathbb{P}(B)$ when $n \to \infty$.

$3' \Longrightarrow 3$: $\bigcup_{i \geq 1} A_i = A_1 \cup A_2 \cup \cdots \cup A_n \cup \Big(\bigcup_{i \geq n} A_i \Big)$.

$\mathbb{P}\Big(\bigcup_{i \geq 1} A_i \Big) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + \mathbb{P}\Big(B_n \Big)$ where $B_n = \bigcup_{i \geq n} A_i$.

Since $B_n \supseteq B_{n+1}$ we have $\mathbb{P}(B_n) \to \mathbb{P}\Big(\bigcap_{n \geq 1} B_n \Big) = \mathbb{P}(\emptyset) = 0$ because A_i's are disjoint.

□

Given algebra A, let $\mathcal{A} = \sigma(A)$ be a σ-algebra generated by A, i.e. intersection of all σ-algebras that contain A. It is easy to see that intersection of all such σ-algebras is itself a σ-algebra. Indeed, consider a sequence A_i for $i \geq 1$ such that each A_i belongs to all σ-algebras that contains A. Then $\bigcup_{i \geq 1} A_i$ belongs to all these σ-algebras and therefore to their intersection.

Let us recall an important result from measure theory.

Theorem 1 *(Caratheodory extension) If A is an algebra of sets and $\mu : A \to \mathbb{R}$ is a non-negative countably additive function on A, then μ can be extended to a measure on σ-algebra $\sigma(A)$. If μ is σ-finite, then this extension is unique. (σ-finite means that $\Omega = \bigcup_{i \geq 1} A_i$ for disjoint sequence A_i and $\mu(A_i) < \infty$.)*

Example. Let A be an algebra of sets $\bigcup_{i \leq n} [a_i, b_i)$ where all $[a_i, b_i)$ are disjoint and $n \geq 1$. Let

$$\lambda\Big(\bigcup_{i \leq n} [a_i, b_i) \Big) = \sum_{i=1}^{n} |b_i - a_i|.$$

One can prove that λ is countably additive on A and therefore can be extended to a Lebesgue measure λ on a sigma algebra $\sigma(A)$ of Borel-measurable sets.

□

Lemma 2 *(Approximation property) If A is an algebra of sets then for any $B \in \sigma(A)$ there exists a sequence $B_n \in A$ such that $\mathbb{P}(B \triangle B_n) \to 0$.*

Remark. Here $\triangle$ denotes symmetric difference $(B \cup B_n) \setminus (B \cap B_n)$. Lemma states that any B in $\sigma(A)$ can be approximated by elements of A.

Proof. Let

$$\mathcal{D} = \{ B \in \sigma(A) : \exists B_n \in A, \mathbb{P}(B \triangle B_n) \to 0 \}.$$

We will prove that $\mathcal{D}$ is a σ-algebra and since $A \subseteq \mathcal{D}$ this will imply that $\sigma(A) \subseteq \mathcal{D}$. One can easily check that

$$d(B, C) := \mathbb{P}(B \triangle C)$$

is a metric. It is also easy to check that

1. $d(BC, DE) \leq d(B, D) + d(C, E)$,
2. $|\mathbb{P}(B) - \mathbb{P}(C)| \leq d(B, C)$,
3. $d(B^c, C^c) = d(B, C)$.

Consider $D_1, \ldots, D_n \in \mathcal{D}$. If a sequence $C_{ij} \in A$ for $j \geq 1$ approximates D_i,

$$\mathbb{P}(C_{ij} \triangle D_i) \to 0, j \to \infty$$

then by properties 1 - 3, $C_j^n := \bigcup_{i \leq n} C_{ij}$ approximates $D^n := \bigcup_{i \leq n} D_i$, which means that $D^n \in \mathcal{D}$. Let $D = \bigcup_{i \geq 1} D_i$. Then

$$\mathbb{P}(D) = \mathbb{P}(D^n) + \mathbb{P}(D \setminus D^n)$$

and obviously $\mathbb{P}(D \setminus D^n) \to 0$ as $n \to \infty$. Therefore, $D \in \mathcal{D}$ and $\mathcal{D}$ is a σ-algebra.

□

Section 2

Random variables and their properties. Expectation.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathcal{S}, \mathcal{B})$ be a measurable space where $\mathcal{B}$ is a σ-algebra of subsets of $\mathcal{S}$. A *random variable* $X : \Omega \to \mathcal{S}$ is a measurable function, i.e.

$$B \in \mathcal{B} \Longrightarrow X^{-1}(B) \in \mathcal{A}.$$

When $\mathcal{S} = \mathbb{R}$ we will usually consider a σ-algebra $\mathcal{B}$ of Borel measurable sets generated by sets $\bigcup_{i \le n}(a_i, b_i]$ (or, equivalently, generated by sets (a_i, b_i) or by open sets).

Lemma 3 *$X : \Omega \to \mathbb{R}$ is a random variable iff for all $t \in \mathbb{R}$*

$$\{X \le t\} := \{\omega \in \Omega : X(\omega) \in (-\infty, t]\} \in \mathcal{A}.$$

Proof. Only $\Leftarrow$ direction requires proof. We will prove that

$$\mathcal{D} = \{D \subseteq \mathbb{R} : X^{-1}(D) \in \mathcal{A}\}$$

is a σ-algebra. Since sets $(-\infty, t] \in \mathcal{D}$ this will imply that $\mathcal{B} \subseteq \mathcal{D}$. The result follows simply because taking pre-image preserves set operations. For example, if we consider a sequence $D_i \in \mathcal{D}$ for $i \ge 1$ then

$$X^{-1}\Big(\bigcup_{i \ge 1} D_i\Big) = \bigcup_{i \ge 1} X^{-1}(D_i) \in \mathcal{A}$$

because $X^{-1}(D_i) \in \mathcal{A}$ and $\mathcal{A}$ is a σ-algebra. Therefore, $\bigcup_{i \ge 1} D_i \in \mathcal{D}$. Other properties can be checked similarly, so $\mathcal{D}$ is a σ-algebra.

□

Let us define a measure $\mathbb{P}_X$ on $\mathcal{B}$ by $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$, i.e. for $B \in \mathcal{B}$,

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P} \circ X^{-1}(B).$$

$(\mathcal{S}, \mathcal{B}, \mathbb{P}_X)$ is called the *sample space* of a random variable X and $\mathbb{P}_X$ is called *the law* of X. Clearly, on this space a random variable $\xi : \mathcal{S} \to \mathcal{S}$ defined by the identity $\xi(s) = s$ has the same law as X.

When $\mathcal{S} = \mathbb{R}$, a function $F(t) = \mathbb{P}(X \le t)$ is called the cumulative distribution function (c.d.f.) of X.

Lemma 4 *F is a c.d.f. of some r.v. X iff*

1. $0 \le F(t) \le 1$,
2. *F is non-decreasing, right-continuous,*

3. $\lim_{t\to-\infty} F(t) = 0$, $\lim_{t\to+\infty} F(t) = 1$.

Proof. The fact that any c.d.f. satisfies properties 1 - 3 is obvious. Let us show that F which satisfies properties 1 - 3 is a c.d.f. of some r.v. X. Consider algebra A consisting of sets $\bigcup_{i\le n}(a_i, b_i]$ for disjoint intervals and for all $n \ge 1$. Let us define a function $\mathbb{P}$ on A by

$$\mathbb{P}\Big(\bigcup_{i\le n}(a_i, b_i]\Big) = \sum_{i\le n}\big(F(a_i) - F(b_i)\big).$$

One can show that $\mathbb{P}$ is countably additive on A. Then, by Caratheodory extension Theorem 1, $\mathbb{P}$ extends uniquely to a measure $\mathbb{P}$ on $\sigma(A) = \mathcal{B}$ - Borel measurable sets. This means that $(\mathbb{R}, \mathcal{B}, \mathbb{P})$ is a probability space and, clearly, random variable $X : \mathbb{R} \to \mathbb{R}$ defined by $X(x) = x$ has c.d.f. $\mathbb{P}(X \le t) = F(t)$. Below we will sometimes abuse the notations and let F denote both c.d.f. and probability measure $\mathbb{P}$.
Alternative proof. Consider a probability space $([0, 1], \mathcal{B}, \lambda)$, where λ is the Lebesgue measure. Define r.v. $X : [0, 1] \to \mathbb{R}$ by the quantile transformation

$$X(t) = \inf\{x \in \mathbb{R}, F(x) \ge t\}.$$

The c.d.f. of X is $\lambda(t : X(t) \le a) = F(a)$ since

$$X(t) \le a \iff \inf\{x : F(x) \ge t\} \le a \iff \exists a_n \to a, F(a_n) \ge t \iff F(a) \ge t.$$

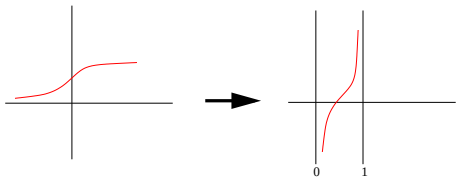


Figure 2.1: A random variable defined by quantile transformation.

□

Definition. Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a r.v. $X : \Omega \to \mathcal{S}$ let $\sigma(X)$ be a σ-algebra generated by a collection of sets $\{X^{-1}(B) : B \in \mathcal{B}\}$. Clearly, $\sigma(X) \subseteq \mathcal{A}$. Moreover, the above collection of sets is itself a σ-algebra. Indeed, consider a sequence $A_i = X^{-1}(B_i)$ for some $B_i \in \mathcal{B}$. Then

$$\bigcup_{i\ge1} A_i = \bigcup_{i\ge1} X^{-1}(B_i) = X^{-1}\Big(\bigcup_{i\ge1} B_i\Big) = X^{-1}(B)$$

where $B \in \bigcup_{i\ge1} B_i \in \mathcal{B}$. $\sigma(X)$ is called the *σ-algebra generated by a r.v. X.*

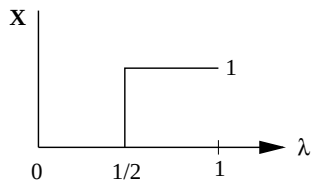


Figure 2.2: $\sigma(X)$ generated by X.

Example. Consider a r.v. defined in figure 2.2. We have $\mathbb{P}(X = 0) = \frac{1}{2}, \mathbb{P}(X = 1) = \frac{1}{2}$ and

$$\sigma(X) = \Big\{\emptyset, \Big[0, \frac{1}{2}\Big], \Big(\frac{1}{2}, 1\Big], [0, 1]\Big\}.$$

Lemma 5 *Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a measurable space $(\mathcal{S}, \mathcal{B})$ and random variables $X : \Omega \to \mathcal{S}$ and $Y : \Omega \to \mathbb{R}$. Then the following are equivalent:*

1. *$Y = g(X)$ for some (Borel) measurable function $g : \mathcal{S} \to \mathbb{R}$.*
2. *$Y : \Omega \to \mathbb{R}$ is measurable on $(\Omega, \sigma(X))$, i.e. with respect to the σ-algebra generated by X.*

Remark. It should be obvious from the proof that $\mathbb{R}$ can be replaced by any separable metric space.

Proof. The fact that 1 implies 2 is obvious since for any Borel set $B \subseteq \mathbb{R}$ the set $B' := g^{-1}(B) \in \mathcal{B}$ and, therefore,

$$\{Y = g(X) \in B\} = \{X \in g^{-1}(B) = B'\} = X^{-1}(B') \in \sigma(X).$$

Let us show that 2 implies 1. For all integer n and k consider sets

$$A_{n,k} = \Big\{\omega : Y(\omega) \in \Big[\frac{k}{2^n}, \frac{k+1}{2^n}\Big)\Big\} = Y^{-1}\Big(\Big[\frac{k}{2^n}, \frac{k+1}{2^n}\Big)\Big).$$

By 2, $A_{n,k} \in \sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\}$ and, therefore, $A_{n,k} = X^{-1}(B_{n,k})$ for some $B_{n,k} \in \mathcal{B}$. Let us consider a function

$$g_n(X) = \sum_{k \in \mathbb{Z}} \frac{k}{2^n}\ \mathrm{I}(X \in B_{n,k}).$$

By construction, $|Y - g_n(X)| \le \frac{1}{2^n}$ since

$$Y(\omega) \in \Big[\frac{k}{2^n}, \frac{k+1}{2^n}\Big) \iff X(\omega) \in B_{n,k} \iff g_n(X(\omega)) = \frac{k}{2^n}.$$

It is easy to see that $g_n(x) \le g_{n+1}(x)$ and, therefore, $g(x) = \lim_{n\to\infty} g_n(x)$ is a measurable function on $(\mathcal{S}, \mathcal{B})$ and, clearly, $Y = g(X)$.

□

Discrete random variables.

A r.v. $X : \Omega \to \mathcal{S}$ is called discrete if $\mathbb{P}_X(\{S_i\}_{i\ge1}) = 1$ for some sequence $S_i \in \mathcal{S}$.

□

Absolutely continuous random variables.

On a measure space $(\mathcal{S}, \mathcal{B})$, a measure $\mathbb{P}$ is called *absolutely continuous* w.r.t. a measure λ if

$$\forall B \in \mathcal{B}, \lambda(B) = 0 \Longrightarrow \mathbb{P}(B) = 0.$$

The following is a well known result from measure theory.

Theorem 2 *(Radon-Nikodym) If $\mathbb{P}$ and λ are sigma-finite and $\mathbb{P}$ is absolutely continuous w.r.t. λ then there exists a* Radon-Nikodym derivative *$f \ge 0$ such that for all $B \in \mathcal{B}$*

$$\mathbb{P}(B) = \int_B f(s) d\lambda(s).$$

f is uniquely defined up to a λ-null sets.

In a typical setting of $\mathcal{S} = \mathbb{R}^k$, a probability measure $\mathbb{P}$ and Lebesgue's measure λ, f is called the *density* of the distribution $\mathbb{P}$.

□

Independence.

Consider a probability space $(\Omega, \mathcal{C}, \mathbb{P})$ and two σ-algebras $\mathcal{A}, \mathcal{B} \subseteq \mathcal{C}$. $\mathcal{A}$ and $\mathcal{B}$ are called *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \ \text{ for all } \ A \in \mathcal{A}, B \in \mathcal{B}.$$

σ-algebras $\mathcal{A}_i \subseteq \mathcal{C}$ for $i \leq n$ are *independent* if

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i \leq n} \mathbb{P}(A_i) \quad \text{for all} \quad A_i \in \mathcal{A}_i.$$

σ-algebras $\mathcal{A}_i \subseteq \mathcal{C}$ for $i \leq n$ are *pairwise independent* if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \quad \text{for all} \quad A_i \in \mathcal{A}_i, A_j \in \mathcal{A}_j, i \neq j.$$

Random variables $X_i : \Omega \to \mathcal{S}$ for $i \leq n$ are (pairwise) independent if σ-algebras $\sigma(X_i), i \leq n$ are (pairwise) independent which is just another convenient way to state the familiar

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \times \ldots \times \mathbb{P}(X_n \in B_n)$$

for any events $B_1, \ldots, B_n \in \mathcal{B}$.

Example. Consider a regular tetrahedron die, Figure 2.3, with red, green and blue sides and a red-green-blue base. If we roll this die then indicators of different colors provide an example of pairwise independent r.v.s that are not independent since

$$\mathbb{P}(r) = \mathbb{P}(b) = \mathbb{P}(g) = \frac{1}{2} \text{ and } \mathbb{P}(rb) = \mathbb{P}(rg) = \mathbb{P}(bg) = \frac{1}{4}$$

but

$$\mathbb{P}(rbg) = \frac{1}{4} \neq \mathbb{P}(r)\mathbb{P}(b)\mathbb{P}(g) = \left(\frac{1}{2}\right)^3.$$

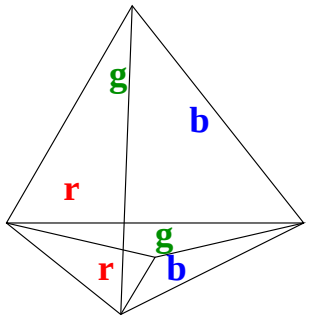


Figure 2.3: Pairwise independent but not independent r.v.s.

□

Independence of σ-algebras can be checked on generating algebras:

Lemma 6 *If algebras $A_i, i \leq n$ are independent then σ-algebras $\sigma(A_i)$ are independent.*

Proof. Obvious by Approximation Lemma 2.

□

Lemma 7 *Consider r.v.s $X_i : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.*

1. *X_i's are independent iff*

$$\mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_n) = \mathbb{P}(X_1 \leq t_1) \times \ldots \times \mathbb{P}(X_n \leq t_n). \tag{2.0.1}$$

2. *If the laws of X_i's have densities $f_i(x)$ then X_i's are independent iff a joint density exists and*

$$f(x_1, ..., x_n) = \prod f_i(x_i).$$

Proof. 1 is obvious by Lemma 6 because (2.0.1) implies the same equality for intervals

$$\mathbb{P}(X_1 \in (a_1, b_1], \ldots, X_n \in (a_n, b_n]) = \mathbb{P}(X_1 \in (a_1, b_1]) \times \ldots \times \mathbb{P}(X_n \in (a_n, b_n])$$

and, therefore, for finite union of disjoint such intervals. To check this for intervals (for example, for $n = 2$) we can write $\mathbb{P}(a_1 < X_1 \le b_1, a_2 < X_n \le b_2)$ as

$$\begin{aligned} & \mathbb{P}(X_1 \le b_1, X_2 \le b_2) - \mathbb{P}(X_1 \le a_1, X_2 \le b_2) - \mathbb{P}(X_1 \le b_1, X_2 \le a_2) + \mathbb{P}(X_1 \le a_1, X_2 \le a_2) \\ = \;& \mathbb{P}(X_1 \le b_1)\mathbb{P}(X_2 \le b_2) - \mathbb{P}(X_1 \le a_1)\mathbb{P}(X_2 \le b_2) - \mathbb{P}(X_1 \le b_1)\mathbb{P}(X_2 \le a_2) + \mathbb{P}(X_1 \le a_1)\mathbb{P}(X_2 \le a_2) \\ = \;& \big(\mathbb{P}(X_1 \le b_1) - \mathbb{P}(X_1 \le a_1)\big)\big(\mathbb{P}(X_2 \le b_2) - \mathbb{P}(X_2 \le a_2)\big) = \mathbb{P}(a_1 < X_1 \le b_1)\mathbb{P}(a_2 < X_2 \le b_2). \end{aligned}$$

To prove 2 we start with "$\Longleftarrow$".

$$\begin{aligned} \mathbb{P}(\cap\{X_i \in A_i\}) &= \mathbb{P}(\boldsymbol{X} \in A_1 \times \cdots \times A_n) = \int_{A_1 \times \cdots \times A_n} \prod f_i(x_i) d\boldsymbol{x} \\ &= \prod \int_{A_i} f_i(x_i) dx_i \ \{\text{by Fubini's Theorem}\} = \prod_{i \le n} \mathbb{P}(X_i \in A_i). \end{aligned}$$

Next, we prove "$\Longrightarrow$". First of all, by independence,

$$\mathbb{P}(\boldsymbol{X} \in A_1 \times \cdots \times A_n) = \prod \mathbb{P}(X_i \in A_i) \overset{\text{Fubini}}{=} \int_{A_1 \times \cdots \times A_n} \prod f_i(x_i) d\boldsymbol{x}.$$

Therefore, the same equality holds for sets in algebra A that consists of finite unions of disjoint sets $A_1 \times \cdots \times A_n$, i.e.

$$\mathbb{P}(\boldsymbol{X} \in B) = \int_B \prod f_i(x_i) d\boldsymbol{x} \text{ for } B \in A.$$

Both $\mathbb{P}(X \in B), \int_B \prod f_i(x_i) d\boldsymbol{x}$ are countably additive on A and finite,

$$\mathbb{P}(\mathbb{R}^n) = \int_{\mathbb{R}^n} \prod f_i(x_i) d\boldsymbol{x} = 1.$$

By the Caratheodory extension Theorem 1, they extend uniquely to all Borel sets $\mathcal{B} = \sigma(A)$, so

$$\mathbb{P}(B) = \int_B \prod f_i(x_i) d\boldsymbol{x} \text{ for } B \in \mathcal{B}.$$

□

Expectation. If $X : \Omega \to \mathbb{R}$ is a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ then *expectation* of X is defined as

$$\mathbb{E}X = \int_\Omega X(\omega) d\mathbb{P}(\omega).$$

In other words, expectation is just another term for the integral with respect to a probability measure and, as a result, expectation has all the usual properties of the integrals. Let us emphasize some of them.

Lemma 8 *1. If F is the c.d.f. of X then for any measurable function $g : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}g(x) = \int_{\mathbb{R}} g(x) dF(x).$$

2. If X is discrete, i.e. $\mathbb{P}(X \in \{x_i\}_{i \ge 1}) = 1$, then

$$\mathbb{E}X = \sum_{i \ge 1} x_i \mathbb{P}(X = x_i).$$

3. If $X : \Omega \to \mathbb{R}^k$ has a density $f(x)$ on $\mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}$ then

$$\mathbb{E}g(X) = \int g(x)f(x)dx.$$

Proof. All these properties follow by making a change of variables $x = X(\omega)$ or $\omega = X^{-1}(x)$, i.e.

$$\mathbb{E}g(X) = \int_\Omega g(X(\omega))d\mathbb{P}(\omega) = \int_{\mathbb{R}} g(x)d\mathbb{P} \circ X^{-1}(x) = \int_{\mathbb{R}} g(x)d\mathbb{P}_X(x),$$

where $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ is the law of X. Another way to see this would be to start with indicator functions of sets $g(x) = \mathrm{I}(x \in B)$ for which

$$\mathbb{E}g(X) = \mathbb{P}(X \in B) = \mathbb{P}_X(B) = \int_{\mathbb{R}} \mathrm{I}(x \in B)d\mathbb{P}_X(x)$$

and, therefore, the same is true for simple step functions

$$g(x) = \sum_{i \geq n} w_i \mathrm{I}(x \in B_i)$$

for disjoint B_i. By approximation, this is true for any measurable functions.

□

Section 3

Kolmogorov's Theorem about consistent distributions.

The notion of a general probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ on this space are rather abstract and often one is really interested in the law $\mathbb{P}_X$ of X on the sample space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$. One can always define a random variable with this law by taking $X : \mathbb{R} \to \mathbb{R}$ to be the identity $X(x) = x$. Similarly, one can define a random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ on $\mathbb{R}^k$ by defining the distribution on the Borel σ-algebra $\mathcal{B}_k$ first. How can we define a distribution on an infinite dimensional space or, in other words, how can we define an infinite family of random variables

$$(X_t)_{t\in T} \in \mathbb{R}^T = \prod_{t\in T} \mathbb{R}_t = \{f : T \to \mathbb{R}\}$$

for some infinite set T? Obviously, there are various ways to do that, for example, we can define explicitly $X_t = \cos(tU)$ for some random variable U. In this section we will consider a typical situation when we start by defining the distribution on any finite subset of coordinates, i.e. for any finite subset $N \subseteq T$ the law $\mathbb{P}_N$ of $(X_t)_{t\in N}$ on the Borel σ-algebra $\mathcal{B}_N$ on $\mathbb{R}^N$ is given. Clearly, these laws must satisfy a natural *consistency assumption*: for any finite subsets $N \subseteq M$ and any Borel set $B \in \mathcal{B}_N$,

$$\mathbb{P}_N(B) = \mathbb{P}_M(B \times \mathbb{R}^{M-N}). \tag{3.0.1}$$

Then the problem is to define a sample space simultaneously for the entire family $(X_t)_{t\in T}$, i.e. we need to define a σ-algebra $\mathcal{A}$ of measurable events in $\mathbb{R}^T$ and a probability measure $\mathbb{P}$ on it that agrees with our finite dimensional distributions $\mathbb{P}_N$. At the very least, $\mathcal{A}$ should contain events expressed in terms of finite number of coordinates, i.e. the following algebra of sets on $\mathbb{R}^T$:

$$A = \{B \times \mathbb{R}^{T-N} : B \in \mathcal{B}_N\}.$$

(It is easy to check that A is an algebra.) A set $B \times \mathbb{R}^{T-N}$ is called a cylinder and B is the base of the cylinder. The probability $\mathbb{P}$ on such sets is of course defined by by

$$\mathbb{P}(B \times \mathbb{R}^{T-N}) = \mathbb{P}_N(B).$$

Notice that, by consistency assumption, $\mathbb{P}$ is well defined. Given two finite subsets $N_1, N_2 \subset T$ and $B_1 \in \mathcal{B}_{N_1}$, the same set can be respresented as

$$B_1 \times \mathbb{R}^{T-N_1} = \left(B_1 \times \mathbb{R}^{(N_1\cup N_2)\setminus N_1}\right) \times \mathbb{R}^{T\setminus(N_1\cup N_2)}.$$

However, by consistency, $\mathbb{P}$ will not depend on the representation. Let $\mathcal{A} = \sigma(A)$ be a σ-algebra generated by algebra A, i.e. the minimal σ-algebra that contains all cylinders.

Definition. A is called the *cylindrical* algebra and $\mathcal{A}$ is the *cylindrical* σ-algebra on $\mathbb{R}^T$.

Example. If $\mathbb{N} \subseteq T$ then $\{\sup_{i\geq 1} X_i \leq 1\}$ is a measurable event in $\mathcal{A}$.

Theorem 3 *(Kolmogorov) For consistent family of distributions (3.0.1), $\mathbb{P}$ can be uniquely extended to $\mathcal{A}$.*

Proof. To use the Caratheodory extension Theorem 1, we need to show that $\mathbb{P}$ is countably additive on A or, equivalently, that it satisfies continuity of measure property: given a sequence $B_n \in A$,

$$B_n \supseteq B_{n+1}, \bigcap_{n\geq 1} B_n = \emptyset \Longrightarrow \mathbb{P}(B_n) \to 0.$$

We will prove that if there exists $\varepsilon > 0$ such that $\mathbb{P}(B_n) > \varepsilon$ for all n then $\bigcap_{n\geq 1} B_n \neq \emptyset$. We have

$$B_n = C_n \times \mathbb{R}^{T-N_n},\ N_n \text{ - finite subset of } T \text{ and } C_n \in \mathcal{B}_{N_n}.$$

Since $B_n \supseteq B_{n+1}$, we can assume that $N_n \subseteq N_{n+1}$. First of all, by regularity of measure $\mathbb{P}_{N_n}$ there exists a compact set $K_n \subseteq C_n$ such that

$$\mathbb{P}_{N_n}(C_n \setminus K_n) \leq \frac{\varepsilon}{2^{n+1}}.$$

We have,

$$\bigcap_{i\leq n} C_i \times \mathbb{R}^{T-N_i} \setminus \bigcap_{i\leq n} K_i \times \mathbb{R}^{T-N_i} \subseteq \bigcup_{i\leq n} (C_i \setminus K_i) \times \mathbb{R}^{T-N_i}$$

and, therefore,

$$\begin{aligned} \mathbb{P}\Big(\bigcap_{i\leq n} C_i \times \mathbb{R}^{T-N_i} \setminus \bigcap_{i\leq n} K_i \times \mathbb{R}^{T-N_i}\Big) &\leq \mathbb{P}\Big(\bigcup_{i\leq n} (C_i \setminus K_i) \times \mathbb{R}^{T-N_i}\Big) \\ &\leq \sum_{i\leq n} \mathbb{P}\Big((C_i \setminus K_i) \times \mathbb{R}^{T-N_i}\Big) \leq \sum_{i\leq n} \frac{\varepsilon}{2^{i+1}} \leq \frac{\varepsilon}{2}. \end{aligned}$$

Since $\mathbb{P}(B_n) = \mathbb{P}\Big(\bigcap_{i\leq n} C_i \times \mathbb{R}^{T-N_i}\Big) > \varepsilon$ this implies that

$$\mathbb{P}\Big(\bigcap_{i\leq n} K_i \times \mathbb{R}^{T-N_i}\Big) \geq \frac{\varepsilon}{2} > 0.$$

We can write

$$\bigcap_{i\leq n} K_i \times \mathbb{R}^{T-N_i} = \bigcap_{i\leq n} (K_i \times \mathbb{R}^{N_n-N_i}) \times \mathbb{R}^{T-N_n} = K^n \times \mathbb{R}^{T-N_n}$$

where $K^n = \bigcap_{i\leq n} (K_i \times \mathbb{R}^{N_n-N_i})$ is a compact in $\mathbb{R}^{N_n}$, since K_n is a compact in $\mathbb{R}^{N_n}$. We proved that

$$\mathbb{P}_{N_n}(K^n) = \mathbb{P}(K^n \times \mathbb{R}^{T-N_n}) = \mathbb{P}\Big(\bigcap_{i\leq n} K_i \times \mathbb{R}^{T-N_i}\Big) > 0$$

and, therefore, there exists a point

$$x^n = (x_1^n, \ldots, x_{N_n}^n, \ldots) \in K^n \times \mathbb{R}^{T-N_n}.$$

We also have the following inclusion property. For $m > n$,

$$x^m \in K^m \times \mathbb{R}^{T-N_m} \subseteq K^n \times \mathbb{R}^{T-N_n}$$

and, therefore, $(x_1^m, ..., x_{N_n}^m) \in K^n$. Any sequence on a compact has a converging subsequence. Let $\{n_k^1\}_{k\geq 1}$ be such that $(x_1^{n_k^1}, \ldots, x_{N_1}^{n_k^1}) \overset{k\to\infty}{\longrightarrow} (x_1, \ldots, x_{N_1}) \in K^1$. Then we can take a subsequence $\{n_k^2\}_{k\geq 1} \subseteq \{n_k^1\}_{k\geq 1}$ such that $(x_1^{n_k^1}, \ldots, x_{N_2}^{n_k^2}) \longrightarrow (x_1, \ldots, x_{N_2}) \in K^2$. By iteration, we can find a subsequence $\{n_k^m\}_{k\geq 1} \subseteq \{n_k^{m-1}\}_{k\geq 1}$, such that

$$(x_1^{n_k^m}, ..., x_{N_m}^{n_k^m}) \longrightarrow (x_1, ..., x_{N_m}) \in K^m.$$

Therefore, a point

$$(x_1, x_2, \ldots) \in \bigcap_{n\geq 1} K^n \times \mathbb{R}^{T-N_n} \subseteq \bigcap_{n\geq 1} B_n,$$

so this last set is not empty.

□

Section 4

Laws of Large Numbers.

Consider a r.v. X and sequence of r.v.s $(X_n)_{n\geq 1}$ on some probability space. We say that X_n converges to X *in probability* if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$

We say that X_n converges to X *almost surely* or *with probability* 1 if

$$\mathbb{P}(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1.$$

Lemma 9 *(Chebyshev's inequality) If a r.v. $X \geq 0$ then for $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}.$$

Proof.

$$\mathbb{E}X = \mathbb{E}X\mathrm{I}(X < t) + \mathbb{E}X\mathrm{I}(X \geq t) \geq \mathbb{E}X\mathrm{I}(X \geq t) \geq t\mathbb{E}\mathrm{I}(X \geq t) = t\mathbb{P}(X \geq t).$$

□

Theorem 4 *(Weak law of large numbers) Consider a sequence of r.v.s $(X_i)_{i\geq 1}$ that are centered, $\mathbb{E}X_i = 0$, have finite second moments, $\mathbb{E}X_i^2 \leq K < \infty$ and are uncorrelated, $\mathbb{E}X_iX_j = 0, i \neq j$. Then*

$$S_n = \frac{1}{n}\sum_{i\leq n} X_i \to 0$$

in probability.

Proof. By Chebyshev's inequality we have

$$\begin{aligned}\mathbb{P}(|S_n - 0| \geq \varepsilon) &= \mathbb{P}(S_n^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}S_n^2}{\varepsilon^2} \\ &= \frac{1}{n^2\varepsilon^2}\mathbb{E}(X_1 + \cdots + X_n)^2 = \frac{1}{n^2\varepsilon^2}\sum_{i\leq n}\mathbb{E}X_i^2 \leq \frac{nK}{n^2\varepsilon^2} = \frac{K}{n\varepsilon^2} \overset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

□

Example. Before we turn to a.s. convergence results, let us note that convergence in probability is weaker than a.s. convergence. For example, consider a probability space which is a circle of circumference 1 with uniform measure on it. Consider a sequence of r.v. on this probability space defined by

$$X_k(x) = \mathrm{I}\Big(x \in \Big[1 + \frac{1}{2} + \cdots + \frac{1}{k}, 1 + \cdots + \frac{1}{k+1}\Big) \bmod 1\Big).$$

Then $X_k \to 0$ in probability, since for $0 < \varepsilon < 1$

$$\mathbb{P}(|X_k - 0| \geq \varepsilon) = \frac{1}{k+1} \to 0$$

but, clearly, X_k does not converge a.s. because the series $\sum_{k\geq 1} 1/k$ diverges and, as a result, each point x on the sphere will fall into the above intervals infinitely many times, i.e. it will satisfy $X_k(x) = 1$ for infinitely many k.

□

Lemma 10 *Consider a sequence $(p_i)_{i\geq 1}$ such that $p_i \in [0, 1)$. Then*

$$\prod_{i\geq 1}(1 - p_i) = 0 \Longleftrightarrow \sum_{i\leq 1} p_i = +\infty.$$

Proof. "$\Longleftarrow$". Using that $1 - p \leq e^{-p}$ we get

$$\prod_{i\leq n}(1 - p_i) \leq \exp(-\sum_{i\leq n} p_i) \to 0 \text{ as } n \to \infty.$$

"$\Longrightarrow$". We can assume that $p_i \leq \frac{1}{2}$ for $i \geq m$ for large enough m because, otherwise, the series obviously diverges. Since $1 - p \geq e^{-2p}$ for $p \leq 1/2$ we have

$$\prod_{m\leq i\leq n}(1 - p_i) \geq \exp\Big(-2 \sum_{m\leq i\leq n} p_i\Big)$$

and the result follows.

□

Lemma 11 *(Borel-Cantelli) Consider a sequence $(A_n)_{n\geq 1}$ of events $A_n \in \mathcal{A}$ on probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Consider an event*

$$A_n \;\; i.o. := \limsup A_n := \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m$$

that A_n occur infinitely often. Then

1. *$\sum_{n\geq 1} \mathbb{P}(A_n) < \infty \Longrightarrow \mathbb{P}(A_n \; i.o.) = 0$.*
2. *If A_n are independent then $\sum_{n\geq 1} \mathbb{P}(A_n) = +\infty \Longrightarrow \mathbb{P}(A_n \; i.o.) = 1$.*

Proof. 1. If $B_n = \bigcup_{m\geq n} A_m$ then $B_n \supseteq B_{n+1}$ and by continuity of measure

$$\mathbb{P}(A_n \text{ i.o}) = \mathbb{P}\Big(\bigcap_{n\geq 1} B_n\Big) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

We have

$$\mathbb{P}(B_n) = \mathbb{P}\Big(\bigcup_{m\geq n} A_m\Big) \leq \sum_{m\geq n} \mathbb{P}(A_m) \to 0 \text{ as } n \to +\infty \text{ because } \sum_{m\geq 1} \mathbb{P}(A_m) < \infty.$$

2. We have

$$\begin{aligned}
\mathbb{P}\Big(\Omega \setminus B_n\Big) &= \mathbb{P}\Big(\Omega \setminus \bigcup_{m\geq n} A_m\Big) = \mathbb{P}\Big(\bigcap_{m\geq n} (\Omega \setminus A_m)\Big) \\
&= \prod_{m\geq n} \mathbb{P}(\Omega \setminus A_m) (\text{by independence}) = \prod_{m\geq n} (1 - \mathbb{P}(A_m)) = 0,
\end{aligned}$$

by Lemma 10, since $\sum_{m\geq n} \mathbb{P}(A_m) = +\infty$. Therefore, $\mathbb{P}(B_n) = 1$ and $\mathbb{P}(A_n \text{ i.o}) = \mathbb{P}\Big(\bigcap_{n\geq 1} B_n\Big) = 1$.

□

Strong law of large numbers. The following simple observation will be useful. If a random variable $X \geq 0$ then $\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq x)dx$. Indeed,

$$\mathbb{E}X = \int_0^\infty x dF(x) = \int_0^\infty \int_0^x 1 ds dF(x) = \int_0^\infty \int_s^\infty 1 dF(x) ds = \int_0^\infty \mathbb{P}(X \geq s) ds.$$

For $X \geq 0$ such that $\mathbb{E}X < \infty$ this implies

$$\sum_{i \geq 1} \mathbb{P}(X \geq i) \leq \int_0^\infty \mathbb{P}(X \geq s) ds = \mathbb{E}X < \infty.$$

Theorem 5 *(Strong law of large numbers) If $\mathbb{E}|X| < \infty$ and $(X_i)_{i \geq 1}$ are i.i.d. copies of X then*

$$\frac{1}{n} \sum_{i=1}^n X_i \to \mathbb{E}X_1 \text{ almost surely (a.s.).}$$

Proof. The proof will proceed in several steps.

1. First, without loss of generality we can assume that $X_i \geq 0$. Indeed, for signed r.v.s we can decompose $X_i = X_i^+ - X_i^-$ where

$$X_i^+ = X_i \mathrm{I}(X_i \geq 0) \text{ and } X_i^- = |X_i| \mathrm{I}(X_i < 0)$$

and the general result would follow since

$$\frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \sum_{i=1}^n X_i^+ - \frac{1}{n} \sum_{i=1}^n X_i^- \to \mathbb{E}X_1^+ - \mathbb{E}X_1^- = \mathbb{E}X_1.$$

Thus, from now on we assume that $X_i \geq 0$.

2. (Truncation) Next, we can replace X_i by $Y_i = X_i \mathrm{I}(X_i \leq i)$ using Borel-Cantelli lemma. We have

$$\sum_{i \geq 1} \mathbb{P}(X_i \neq Y_i) = \sum_{i \geq 1} \mathbb{P}(X_i > i) \leq \mathbb{E}X_1 < \infty$$

and Borel-Cantelli lemma implies that $\mathbb{P}(\{X_i \neq Y_i\} \text{ i.o.}) = 0$. This means that for some (random) i_0 and for $i \geq i_0$ we have $X_i = Y_i$ and, therefore,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n X_i = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n Y_i.$$

It remains to show that if $T_n = \sum_{i=1}^n Y_i$ then $\frac{T_n}{n} \to \mathbb{E}X$ a.s.

3. (Limit over subsequences) We will first prove this along the subsequences $n(k) = \lfloor \alpha^k \rfloor$ for $\alpha > 1$. For any $\varepsilon > 0$,

$$\begin{aligned}
\sum_{k \geq 1} \mathbb{P}\Big(|T_{n(k)} - \mathbb{E}T_{n(k)}| \geq \varepsilon n(k) \Big) &\leq \sum_{k \geq 1} \frac{1}{\varepsilon^2 n(k)^2} \mathrm{Var}(T_{n(k)}) = \sum_{k \geq 1} \frac{1}{\varepsilon^2 n(k)^2} \sum_{i \leq n(k)} \mathrm{Var}(Y_i) \\
&\leq \sum_{k \geq 1} \frac{1}{\varepsilon^2 n(k)^2} \sum_{i \leq n(k)} \mathbb{E}Y_i^2 = \frac{1}{\varepsilon^2} \sum_{i \geq 1} \mathbb{E}Y_i^2 \sum_{k : n(k) \geq i} \frac{1}{n(k)^2} \\
&\overset{(*)}{\leq} K \sum_{i \geq 1} \mathbb{E}Y_i^2 \frac{1}{i^2} = K \sum_{i \geq 1} \frac{1}{i^2} \int_0^i x^2 dF(x) \leq K \times (**)
\end{aligned}$$

where $F(x)$ is the law of X and a constant $K = K(\alpha)$ depends only on α. (*) follows from

$$\frac{\alpha^k}{2} \leq n(k) = \lfloor \alpha^k \rfloor \leq \alpha^k$$

and if $k_0 = \min\{k : \alpha^k \geq i\}$ then

$$\sum_{n(k)\geq i} \frac{1}{n(k)^2} \leq \sum_{\alpha^k \geq i} \frac{4}{\alpha^{2k}} = \frac{4}{\alpha^{2k_0}(1-\frac{1}{\alpha^2})} \leq \frac{K}{i^2}.$$

We can continue,

$$\begin{aligned}(**) &= \sum_{i\geq 1} \frac{1}{i^2} \sum_{m<i} \int_m^{m+1} x^2 dF(x) = \sum_{m\geq 0} \sum_{i>m} \frac{1}{i^2} \int_m^{m+1} x^2 dF(x) \\ &\leq \sum_{m\geq 0} \frac{1}{m+1} \int_m^{m+1} x^2 dF(x) \leq \sum_{m\geq 0} \int_m^{m+1} x dF(x) = \mathbb{E}X < \infty.\end{aligned}$$

Thus, we proved that

$$\sum_{k\geq 1} \mathbb{P}\Big(|T_{n(k)} - \mathbb{E}T_{n(k)}| \geq \varepsilon n(k)\Big) < \infty$$

and Borel-Cantelli lemma implies that

$$\mathbb{P}\Big(|T_{n(k)} - \mathbb{E}T_{n(k)}| \geq \varepsilon n(k) \text{ i.o.}\Big) = 0.$$

This means that for some (random) k_0

$$\mathbb{P}\Big(\forall k \geq k_0, |T_{n(k)} - \mathbb{E}T_{n(k)}| \leq \varepsilon n(k)\Big) = 1.$$

If we take a sequence $\varepsilon_m = \frac{1}{m}$, this implies that

$$\mathbb{P}\Big(\forall m \geq 1, k \geq k_0(m), |T_{n(k)} - \mathbb{E}T_{n(k)}| \leq \frac{1}{m} n(k)\Big) = 1$$

and this proves that

$$\frac{T_{n(k)}}{n(k)} - \frac{\mathbb{E}T_{n(k)}}{n(k)} \to 0 \text{ a.s.}$$

On the other hand,

$$\frac{1}{n(k)}\mathbb{E}T_{n(k)} = \frac{1}{n(k)} \sum_{i\leq n(k)} \mathbb{E}X_i \mathrm{I}(X_i \leq i) \to \mathbb{E}X \text{ as } k \to \infty,$$

by Lebesgue's dominated convergence theorem. We proved that

$$\frac{T_{n(k)}}{n(k)} \to \mathbb{E}X \text{ a.s.}$$

4. Finally, for j such that

$$n(k) \leq j < n(k+1) = n(k)\frac{n(k+1)}{n(k)} \leq n(k)\alpha^2$$

we can write

$$\frac{1}{\alpha^2}\frac{T_{n(k)}}{n(k)} \leq \frac{T_j}{j} \leq \alpha^2 \frac{T_{n(k+1)}}{n(k+1)}$$

and, therefore,

$$\frac{1}{\alpha^2}\mathbb{E}X \leq \liminf \frac{T_j}{j} \leq \limsup \frac{T_j}{j} \leq \alpha^2 \mathbb{E}X \text{ a.s.}$$

Taking $\alpha = 1 + m^{-1}$ and letting $m \to \infty$ proves that $\lim_{j\to\infty} \frac{T_j}{j} = \mathbb{E}X$ a.s.

□

Section 5

Bernstein Polynomials. Hausdorff and de Finetti theorems.

Let us look at some applications related to the law of large numbers. Consider an i.i.d. sequence of real valued r.v. (X_i) with distribution $\mathbb{P}_\theta$ from a family of distributions parametrized by $\theta \in \Theta \subseteq \mathbb{R}$ such that

$$\mathbb{E}_\theta X_i = \theta, \sigma^2(\theta) := \mathrm{Var}_\theta(X_i) \le K < +\infty.$$

Let $\bar{X}_n = \frac{1}{n}\sum_{i\le n} X_i$. The following holds.

Theorem 6 *If $u : \mathbb{R} \to \mathbb{R}$ is uniformly continuous and bounded then $\mathbb{E}_\theta u(\bar{X}_n) \to u(\theta)$ uniformly over Θ.*

Proof. For any $\varepsilon > 0$,

$$\begin{aligned}
|\mathbb{E}_\theta u(\bar{X}_n) - u(\theta)| &\le \mathbb{E}_\theta |u(\bar{X}_n) - u(\theta)| \\
&= \mathbb{E}_\theta |u(\bar{X}_n) - u(\theta)| \Big(\mathrm{I}(|\bar{X}_n - \theta| \le \varepsilon) + \mathrm{I}(|\bar{X}_n - \theta| > \varepsilon) \Big) \\
&\le \max_{|x-\theta|\le\varepsilon} |u(x) - u(\theta)| + 2 \max_x |u(x)| \mathbb{P}_\theta(|\bar{X}_n - \theta| > \varepsilon) \\
&\le \delta(\varepsilon) + 2\|u\|_\infty \frac{1}{\varepsilon^2} \mathbb{E}_\theta (\bar{X}_n - \theta)^2 \le \delta(\varepsilon) + \frac{2\|u\|_\infty K}{n\varepsilon^2},
\end{aligned}$$

where $\delta(\varepsilon)$ is the modulus of continuity of u. Letting $\varepsilon = \varepsilon_n \to 0$ so that $n\varepsilon_n^2 \to \infty$ finishes the proof. □

Example. Let (X_i) be i.i.d. with Bernoulli distribution $B(\theta)$ with probability of success $\theta \in [0,1]$, i.e.

$$\mathbb{P}_\theta(X_i = 1) = \theta, \;\; \mathbb{P}_\theta(X_i = 0) = 1 - \theta,$$

and let $u : [0,1] \to \mathbb{R}$ be continuous. Then, by the above Theorem, the following *Bernstein polynomials*

$$B_n(\theta) := \mathbb{E}_\theta u(\bar{X}_n) = \sum_{k=0}^n u\Big(\frac{k}{n}\Big) \mathbb{P}_\theta\Big(\sum_{i=1}^n X_i = k\Big) = \sum_{k=0}^n u\Big(\frac{k}{n}\Big) \binom{n}{k} \theta^k (1-\theta)^{n-k} \to u(\theta)$$

uniformly on $[0,1]$.

Example. Let (X_i) have Poisson distribution $\Pi(\theta)$ with intensity parameter $\theta > 0$ defined by

$$\mathbb{P}_\theta(X_i = k) = \frac{\theta^k}{k!} e^{-\theta} \text{ for integer } k \ge 0.$$

Then it is well known (and easy to check) that $\mathbb{E}_\theta X_i = \theta, \sigma^2(\theta) = \theta$ and the sum $X_1 + \ldots + X_n$ has Poisson distribution $\Pi(n\theta)$. If u is bounded and continuous on $[0, +\infty)$ then

$$\mathbb{E}_\theta u(\bar{X}_n) = \sum_{k=0}^\infty u\Big(\frac{k}{n}\Big) \mathbb{P}_\theta\Big(\sum_{i=1}^n X_i = k\Big) = \sum_{k=0}^\infty u\Big(\frac{k}{n}\Big) \frac{(n\theta)^k}{k!} e^{-n\theta} \to u(\theta)$$

uniformly on compact sets.

□

Moment problem. Consider a random variable $X \in [0,1]$ and let $\mu_k = \mathbb{E}X^k$ be its moments. Given a sequence $(c_0, c_1, c_2, \ldots)$ let us define a sequence of increments by $\Delta c_k = c_{k+1} - c_k$. Then

$$-\Delta\mu_k = \mu_k - \mu_{k+1} = \mathbb{E}(X^k - X^{k+1}) = \mathbb{E}X^k(1-X),$$

$$(-\Delta)(-\Delta\mu_k) = (-1)^2\Delta^2\mu_k = \mathbb{E}X^k(1-X) - \mathbb{E}X^{k+1}(1-X) = \mathbb{E}X^k(1-X)^2$$

and by induction

$$(-1)^r\Delta^r\mu_k = \mathbb{E}X^k(1-X)^r.$$

Clearly, $(-1)^r\Delta^r\mu_k \geq 0$ since $X \in [0,1]$. If u is a continuous function on $[0,1]$ and B_n is its corresponding Bernstein polynomial then

$$\mathbb{E}B_n(X) = \sum_{k=0}^n u\Big(\frac{k}{n}\Big)\binom{n}{k}\mathbb{E}X^k(1-X)^{n-k} = \sum_{k=0}^n u\Big(\frac{k}{n}\Big)\binom{n}{k}(-1)^{n-k}\Delta^{n-k}\mu_k.$$

Since $B_n(X)$ converges uniformly to $u(X)$, $\mathbb{E}B_n(X)$ converges to $\mathbb{E}u(X)$. Let us define

$$p_k^{(n)} = \binom{n}{k}(-1)^{n-k}\Delta^{n-k}\mu_k \geq 0, \ \sum_{k=0}^n p_k^{(n)} = 1 \text{ (take } u = 1).$$

We can think of $p_k^{(n)}$ as the distribution of a r.v. $X^{(n)}$ such that

$$\mathbb{P}\Big(X^{(n)} = \frac{k}{n}\Big) = p_k^{(n)}. \tag{5.0.1}$$

We showed that

$$\mathbb{E}B_n(X) = \mathbb{E}u\big(X^{(n)}\big) \to \mathbb{E}u(X)$$

for any continuous function u. We will later see that by definition this means that $X^{(n)}$ converges to X in distribution. Given the moments of a r.v. X, this construction allows us to approximate the distribution of X and expectation of $u(X)$.

□

Next, given a sequence (μ_k), when is it the sequence of moments of some $[0,1]$ valued r.v. X? By the above, it is necessary that

$$\mu_k \geq 0, \mu_0 = 1 \text{ and } (-1)^r\Delta^r\mu_k \geq 0 \text{ for all } k, r. \tag{5.0.2}$$

It turns out that this is also sufficient.

Theorem 7 *(Hausdorff) There exists a r.v. $X \in [0,1]$ such that $\mu_k = \mathbb{E}X^k$ iff (5.0.2) holds.*

Proof. The idea of the proof is as follows. If μ_k are the moments of the distribution of some r.v. X, then the discrete distributions defined in (5.0.1) should approximate it. Therefore, our goal will be to show that condition (5.0.2) ensures that $(p_k^{(n)})$ is indeed a distribution and then show that the moments of (5.0.1) converge to μ_k. As a result, any limit of these distributions will be a candidate for the distribution of X.

First of all, let us express μ_k in terms of $(p_k^{(n)})$. Since $\Delta\mu_k = \mu_{k+1} - \mu_k$ we have the following inversion formula:

$$\begin{aligned}\mu_k &= \mu_{k+1} - \Delta\mu_k = (\mu_{k+2} - \Delta\mu_{k+1}) + (-\Delta\mu_{k+1} + \Delta^2\mu_k) \\ &= \mu_{k+2} - 2\Delta\mu_{k+1} + \Delta^2\mu_k = \sum_{j=0}^r \binom{r}{j}(-1)^{r-j}\Delta^{r-j}\mu_{k+j},\end{aligned}$$

by induction. Take $r = n - k$. Then

$$\mu_k = \sum_{j=0}^{n-k} \frac{\binom{n-k}{j}}{\binom{n}{k+j}} \binom{n}{k+j} (-1)^{n-(k+j)} \Delta^{n-(k+j)} \mu_{k+j} = \sum_{j=0}^{n-k} \frac{\binom{n-k}{j}}{\binom{n}{k+j}} p_{k+j}^{(n)}.$$

We have

$$\frac{\binom{n-k}{j}}{\binom{n}{k+j}} = \frac{(n-k)!}{j!(n-k-j)!} \frac{(k+j)!(n-k-j)!}{n!} = \frac{\binom{k+j}{k}}{\binom{n}{k}}$$

so that

$$\mu_k = \sum_{j=0}^{n-k} \frac{\binom{k+j}{k}}{\binom{n}{k}} p_{k+j}^{(n)} = \sum_{m=k}^{n} \frac{\binom{m}{k}}{\binom{n}{k}} p_m^{(n)}.$$

By (5.0.2), $p_m^{(n)} \geq 0$ and $\sum_{m \leq n} p_m^{(n)} = \mu_0 = 1$ so we can consider a r.v. $X^{(n)}$ such that

$$\mathbb{P}\Big(X^{(n)} = \frac{m}{n}\Big) = p_m^{(n)} \text{ for } 0 \leq m \leq n.$$

We have

$$\begin{aligned} \mu_k &= \sum_{m=k}^{n} \frac{\binom{m}{k}}{\binom{n}{k}} p_m^{(n)} = \sum_{m=k}^{n} \frac{m(m-1)\cdots(m-k+1)}{n(n-1)\cdots(n-k+1)} p_m^{(n)} = \sum_{m=k}^{n} \frac{\frac{m}{n}(\frac{m}{n} - \frac{1}{n})\cdots(\frac{m}{n} - \frac{k+1}{n})}{1(1-\frac{1}{n})\cdots(1-\frac{k+1}{n})} p_m^{(n)} \\ &\overset{n\to\infty}{\approx} \sum_{m=0}^{n} \Big(\frac{m}{n}\Big)^k p_m^{(k)} = \mathbb{E}\Big(X^{(n)}\Big)^k \overset{n\to\infty}{\longrightarrow} \mu_k. \end{aligned}$$

Any continuous function u can be approximated by (for example, Bernstein) polynomials so the limit $\lim_{n\to\infty} \mathbb{E}u\big(X^{(n)}\big)$ exists. By selection theorem that we will prove later in the course, one can choose a subsequence $X^{(n_i)}$ that converges to some r.v. X in distribution and, as a result,

$$\mathbb{E}\Big(X^{(n_i)}\Big)^k \to \mathbb{E}X^k = \mu_k,$$

which means that μ_k are the moments of X.

□

de Finetti's theorem. Consider an *exchangeable* sequence $X_1, X_2, \ldots, X_n, \ldots$ of Bernoulli random variables which means that for any $n \geq 1$ the probability

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n)$$

depends only on $x_1 + \ldots + x_n$, i.e. it does not depend on the order of 1's or 0's. Another way to say this is that for any $n \geq 1$ and any permutation π of $1, \ldots, n$ the distribution of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ does not depend on π. Then the following holds.

Theorem 8 *(de Finetti) There exists a distribution F on $[0,1]$ such that*

$$p_k := \mathbb{P}(X_1 + \ldots + X_n = k) = \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dF(x).$$

This means that to generate such exchangeable sequence we can first pick $x \in [0,1]$ from distribution F and then generate a sequence of i.i.d Bernoulli random variables with probability of success x.

Proof. Let $\mu_0 = 1$ and for $k \geq 1$ define

$$\mu_k = \mathbb{P}(X_1 = 1, ..., X_k = 1). \tag{5.0.3}$$

We have

$$\begin{aligned}\mathbb{P}(X_1=1,...,X_k=1,X_{k+1}=0) &= \mathbb{P}(X_1=1,...,X_k=1)\\ &- \mathbb{P}(X_1=1,...,X_k=1,X_{k+1}=1)\\ &= \mu_k-\mu_{k+1}=-\Delta\mu_k.\end{aligned}$$

Next, using exchangeability

$$\begin{aligned}\mathbb{P}(X_1=1,...,X_k=1,X_{k+1}=0,X_{k+2}=0) &= \mathbb{P}(X_1=1,...,X_k=1,X_{k+1}=0)\\ &- \mathbb{P}(X_1=1,...,X_k=1,X_{k+1}=0,X_{k+2}=1)\\ &= -\Delta\mu_k-(-\Delta\mu_{k+1})=\Delta^2\mu_k.\end{aligned}$$

Similarly, by induction,

$$\mathbb{P}(X_1=1,...,X_k=1,X_{k+1}=0,...,X_n=0)=(-1)^{n-k}\Delta^{n-k}\mu_k\geq 0.$$

By the Hausdorff theorem, $\mu_k=\mathbb{E}X^k$ for some r.v. $X\in[0,1]$ and, therefore,

$$\begin{aligned}\mathbb{P}(X_1=1,...,X_k=1,X_{k+1}=0,...,X_n=0) &= (-1)^{n-k}\Delta^{n-k}\mu_k\\ &= \mathbb{E}X^k(1-X)^{n-k}=\int_0^1 x^k(1-x)^{n-k}dF(x).\end{aligned}$$

Since, by exchangeability, changing the order of 1's and 0's does not affect the probability, we get

$$\mathbb{P}(X_1+\ldots+X_n=k)=\int_0^1\binom{n}{k}x^k(1-x)^{n-k}dF(x).$$

□

Example. (Polya urn model). Suppose we have b blue and r red balls in the urn. We pick a ball

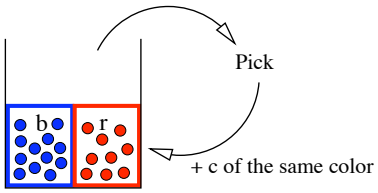


Figure 5.1: Polya urn model.

randomly and return it with c balls of the same color. Consider r.v.s

$$X_i=\begin{cases}1 & \text{if the } i\text{th ball picked is blue}\\ 0 & \text{otherwise.}\end{cases}$$

X_i's are not independent but exchangeable. For example,

$$\mathbb{P}(bbr)=\frac{b}{b+r}\times\frac{b+c}{b+r+c}\times\frac{r}{b+r+2c},\ \ \mathbb{P}(brb)=\frac{b}{b+r}\times\frac{r}{b+r+c}\times\frac{b+r}{b+r+2c}$$

are equal. To identify the distribution F in de Finetti's theorem, let us look at its moments μ_k in (5.0.3),

$$\mu_k=\mathbb{P}(\underbrace{b\ldots b}_{k\text{ times}})=\frac{b}{b+r}\times\frac{b+c}{b+r+c}\times\cdots\times\frac{b+(k-1)c}{b+r+(k-1)c}.$$

One can recognize or easily check that μ_k are the moments of Beta(α,β) distribution with the density

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

on $[0,1]$ with parameters $\alpha = b/c, \beta = r/c$. By de Finetti's theorem, we can generate X_i's by first picking x from distribution Beta$\big(b/c, r/c\big)$ and then generating i.i.d. Bernoulli (X_i)'s with probability of success x. By strong law of large numbers, the proportion of blue balls in the first n repetitions will converge to this probability of success x, i.e. in the limit it will be random with Beta distribution. This example will come up once more when we talk about convergence of martingales.

□

Section 6

0 - 1 Laws. Convergence of random series.

Consider a sequence $(X_i)_{i\geq 1}$ of real valued *independent* random variables and let $\sigma\big((X_i)_{i\geq 1}\big)$ be a σ-algebra of events generated by this sequence, i.e. $\{(X_i)_{i\geq 1} \in B\}$ for B in the cylindrical σ-algebra on $\mathbb{R}^{\mathbb{N}}$.

Definition. An event $A \in \sigma\big((X_i)_{i\geq 1}\big)$ is called a *tail event* if $A \in \sigma\big((X_i)_{i\geq n}\big)$ for all $n \geq 1$.

For example, if $A_i \in \sigma(X_i)$ then

$$A_i \text{ i.o.} = \bigcap_{n\geq 1} \bigcup_{i\geq n} A_i$$

is a tail event. It turns out that such events have probability 0 or 1.

Theorem 9 *(Kolmogorov's 0-1 law) If A is a tail event then $\mathbb{P}(A) = 0$ or 1.*

Proof. For a finite subset $F = \{i_1, \ldots, i_n\} \subset \mathbb{N}$, let us denote by $X_F = (X_{i_1}, \ldots, X_{i_n})$. A σ-algebra $\sigma\big((X_i)_{i\geq 1}\big)$ is generated by algebra

$$\{X_F \in B : F\text{- finite} \subseteq \mathbb{N}, B \in \mathcal{B}(\mathbb{R}^{|F|})\}.$$

By approximation lemma, we can approximate any event $A \in \sigma\big((X_i)_{i\geq 1}\big)$ by events in this generating algebra. Therefore, for any $\varepsilon > 0$ there exists a set A' in this algebra such that $\mathbb{P}(A\Delta A') \leq \varepsilon$ and by definition $A' \in \sigma(X_1, ..., X_n)$ for large enough n. This implies

$$|\mathbb{P}(A) - \mathbb{P}(A')| \leq \varepsilon, \;\; |\mathbb{P}(A) - \mathbb{P}(AA')| \leq \varepsilon.$$

Since A is a tail event, $A \in \sigma((X_i)_{i\geq n+1})$ which means that A, A' are independent, i.e. $\mathbb{P}(AA') = \mathbb{P}(A)\mathbb{P}(A')$. We get

$$\mathbb{P}(A) \approx \mathbb{P}(AA') = \mathbb{P}(A)\mathbb{P}(A') \approx \mathbb{P}(A)\mathbb{P}(A)$$

and letting $\varepsilon \to 0$ proves that $\mathbb{P}(A) = \mathbb{P}(A)^2$.

□

Examples.

1. $\Big\{\sum_{i\geq 1} X_i - \text{ converges}\Big\}$ is a tail event, it has probability 0 or 1.
2. Consider series $\sum_{i\geq 1} X_i z^i$ on a complex plane, $z \in \mathbb{C}$. Its radius of convergence is

$$r = \liminf_{i\to\infty} |X_i|^{-\frac{1}{i}}.$$

For any $x \geq 0$, event $\{r \leq x\}$ is, obviously, a tail event. This implies that $r = \text{const}$ with probability 1.

□

The Savage-Hewitt 0 - 1 law.

Next we will prove a stronger result under more restrictive assumption that the r.v.s $X_i, i \geq 1$ are not only independent but also identically distributed with the law μ. Without loss of generality, we can assume that each X_i is given by the identity $X_i(x) = x$ on its sample space $(\mathbb{R}, \mathcal{B}, \mu)$. By Kolmogorov's consistency theorem the entire sequence $(X_i)_{i\geq 1}$ can be defined on the sample space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\infty}, \mathbb{P})$ where $\mathcal{B}^{\infty}$ is the cylindrical σ-algebra and $\mathbb{P}$ is the measure guaranteed by the Caratheodory extension theorem. In our case X_i's are i.i.d. and $\mathbb{P} = \mu^{\infty}$ is called the infinite product measure. It will be convenient to use the notation $\sigma((X_i)_{i\geq 1})$ for the cylindrical σ-algebra since similar notation can be used for the cylindrical σ-algebra on any subset of coordinates.

Definition. An event $A \in \sigma((X_i)_{i\geq 1})$ - is called *exchangeable/symmetric* if for all $n \geq 1$,

$$(x_1, x_2, \ldots, x_n, x_{n+1}, \ldots) \in A \Longrightarrow (x_n, x_2, ..., x_{n-1}, x_1, x_{n+1}, ...) \in A.$$

In other words, the set A is symmetric under permutations of a finite number of coordinates. Note that any tail event is symmetric.

Theorem 10 *(Savage-Hewitt 0-1 law) If A is symmetric then $\mathbb{P}(A) = 0$ or 1.*

Proof. Given a sequence $x = (x_1, x_2, \ldots)$ let us define an operator

$$\Gamma x = (x_{n+1}, \ldots, x_{2n}, x_1, \ldots, x_n, x_{2n+1}, \ldots)$$

that switches the first n coordinates with the second n coordinates. Since A is symmetric,

$$\Gamma A = \{\Gamma x : x \in A\} = A.$$

By the Approximation Lemma 2 for any $\varepsilon > 0$ for large enough n, there exists $A_n \in \sigma(X_1, ..., X_n)$ such that $\mathbb{P}(A_n \Delta A) \leq \varepsilon$. Clearly,

$$B_n = \Gamma A_n \in \sigma(X_{n+1}, ..., X_{2n})$$

and

$$\mathbb{P}(B_n \Delta A) = \mathbb{P}(\Gamma A_n \Delta \Gamma A) \overset{\text{by i.i.d.}}{=} \mathbb{P}(A_n \Delta A) \leq \varepsilon,$$

which implies that $\mathbb{P}\big((A_n B_n)\Delta A\big) \leq 2\varepsilon$. Therefore, we can conclude that

$$\mathbb{P}(A) \approx \mathbb{P}(A_n), \;\; \mathbb{P}(A) \approx \mathbb{P}(A_n B_n) = \mathbb{P}(A_n)\mathbb{P}(B_n) = \mathbb{P}(A_n)^2$$

where we used the fact that the events A_n, B_n are defined in terms of different sets of coordinates and, thus, are independent. Letting $\varepsilon \to 0$ implies that $\mathbb{P}(A) = \mathbb{P}(A)^2$.

□

Example. Let $S_n = X_1 + \ldots + X_n$ and let

$$r = \limsup_{n\to\infty} \frac{S_n - a_n}{b_n}.$$

Event $\{r \leq x\}$ is symmetric since changing the order of any finite set of coordinates does not affect S_n for large enough n. As a result, $\mathbb{P}(r \leq x) = 0$ or 1, which implies that $r =$ const with probability 1.

□

Random series. We already saw above that, by Kolmogorov's 0-1 law, the series $\sum_{i\geq 1} X_i$ for independent $(X_i)_{i\geq 1}$ converges with probability 0 or 1. This means that either $S_n = X_1 + \ldots + X_n$ converges to its limit S with probability one, or with probability one it does not converge. Two section back, before the proof of the strong law of large numbers, we saw the example of a sequence which with probability one does not converge yet converges to 0 in probability. In case when with probability one S_n does not converge, is it still possible that it converges to some random variable in probability? The answer is no because we will now prove that for random series convergence in probability implies a.s. convergence.

Theorem 11 *(Kolmogorov's inequality) Suppose that $(X_i)_{i\geq 1}$ are independent and $S_n = X_1 + \ldots + X_n$. If for all $j \leq n$,*

$$\mathbb{P}(|S_n - S_j| \geq a) \leq p < 1 \tag{6.0.1}$$

then for $x > a$,

$$\mathbb{P}\Big(\max_{1\leq j\leq n} |S_j| \geq x\Big) \leq \frac{1}{1-p}\mathbb{P}(|S_n| > x - a).$$

Proof. First of all, let us notice that this inequality is obvious without the maximum because (6.0.1) is equivalent to $1 - p \leq \mathbb{P}(|S_n - S_j| < a)$ and we can write

$$\begin{aligned}(1-p)\mathbb{P}\big(|S_j| \geq x\big) &\leq \mathbb{P}\big(|S_n - S_j| < a\big)\mathbb{P}\big(|S_j| \geq x\big) \\ &= \mathbb{P}\big(|S_n - S_j| < a, |S_j| \geq x\big) \leq \mathbb{P}(|S_n| > x - a).\end{aligned}$$

The equality is true because events $\{|S_j| \geq x\}$ and $\{|S_n - S_j| < a\}$ are independent since the first depends only on $X_1, ..., X_j$ and the second only on $X_{j+1}, ..., X_n$. The last inequality is true simply by triangle inequality. To deal with the maximum, instead of looking at an arbitrary partial sum S_j we will look at the first partial sum that crosses level x. We define that first time by $\tau = \min\{j \leq n : |S_j| \geq x\}$ and let $\tau = n + 1$ if all $|S_j| < x$. Notice that event $\{\tau = j\}$ also depends only on $X_1, ..., X_j$ so we can again write

$$\begin{aligned}(1-p)\mathbb{P}\big(\tau = j\big) &\leq \mathbb{P}\big(|S_n - S_j| < a\big)\mathbb{P}\big(\tau = j\big) \\ &= \mathbb{P}\big(|S_n - S_j| < a, \tau = j\big) \leq \mathbb{P}(|S_n| > x - a, \tau = j).\end{aligned}$$

The last inequality is again true by triangle inequality because when $\tau = j$ we have $|S_j| \geq x$ and

$$\Big\{|S_n - S_j| < a, \tau = j\Big\} \subseteq \Big\{|S_n| > x - a, \tau = j\Big\}.$$

It remains to add up over $j \leq n$ to get

$$(1-p)\mathbb{P}(\tau \leq n) \leq \mathbb{P}(|S_n| > x - a, \tau \leq n) \leq \mathbb{P}(|S_n| > x - a)$$

and notice that $\tau \leq n$ is equivalent to $\max_{j\leq n} |S_j| \geq x$.

□

Theorem 12 *(Kolmogorov) If the series $\sum_{i\geq 1} X_i$ converges in probability then it converges almost surely.*

Proof. Suppose that partial sums S_n converge to some r.v. S in probability, i.e. for any $\varepsilon > 0$, for large enough $n \geq n_0(\varepsilon)$ we have $\mathbb{P}(|S_n - S| \geq \varepsilon) \leq \varepsilon$. If $k \geq j \geq n \geq n_0(\varepsilon)$ then

$$\mathbb{P}(|S_k - S_j| \geq 2\varepsilon) \leq \mathbb{P}(|S_k - S| \geq \varepsilon) + \mathbb{P}(|S_j - S| \geq \varepsilon) \leq 2\varepsilon.$$

Next, we use Kolmogorov's inequality for $x = 4\varepsilon$ and $a = 2\varepsilon$ (we let partial sums start at n):

$$\mathbb{P}\Big(\max_{n\leq j\leq k} |S_j - S_n| \geq 4\varepsilon\Big) \leq \frac{1}{1-2\varepsilon}\mathbb{P}(|S_k - S_n| \geq 2\varepsilon) \leq \frac{2\varepsilon}{1-2\varepsilon} \leq 3\varepsilon,$$

for small ε. The events $\{\max_{n\leq j\leq k} |S_j - S_n| \geq 4\varepsilon\}$ are increasing as $k \uparrow \infty$ and by continuity of measure

$$\mathbb{P}\Big(\max_{n\leq j} |S_j - S_n| \geq 4\varepsilon\Big) \leq 3\varepsilon.$$

Finally, since $\mathbb{P}(|S_n - S| \geq \varepsilon) \leq \varepsilon$ we get

$$\mathbb{P}\Big(\max_{n\leq j} |S_j - S| \geq 5\varepsilon\Big) \leq 4\varepsilon.$$

This kind of "maximal" statement about any sequence S_j is actually equivalent to its a.s. convergence. To see this take $\varepsilon = \frac{1}{m^2}$, take $n(m) = n_0(\varepsilon)$ and consider an event

$$A_m = \left\{ \max_{n(m) \le j} |S_j - S| \ge \frac{5}{m^2} \right\}.$$

We proved that

$$\sum \mathbb{P}(A_m) \le \sum \frac{4}{m^2} < \infty$$

and by the Borel-Cantelli lemma, $\mathbb{P}(A_m \text{ i.o.}) = 0$. This means that with probability 1 for large enough (random) m,

$$\max_{j \ge n(m)} |S_j - S| \le \frac{5}{m^2}$$

and, therefore, $S_n \to S$ a.s.

□

Let us give one easy-to-check criterion for convergence of random series, which is called Kolmogorov's strong law of large numbers.

Theorem 13 *If $(X_i)_{i \ge 1}$ is a sequence of independent random variables such that*

$$\mathbb{E}X_i = 0 \text{ and } \sum_{i \ge 1} \mathbb{E}X_i^2 < \infty$$

then $\sum_{i \ge 1} X_i$ converges a.s.

Proof. It is enough to prove convergence in probability. Let us first show the existence of a limit of partial sums S_n over some subsequence. For $m < n$,

$$\mathbb{P}(|S_n - S_m| \ge \varepsilon) \le \frac{1}{\varepsilon^2} \mathbb{E}(S_n - S_m)^2 = \frac{1}{\varepsilon^2} \sum_{m < i \le n} \mathbb{E}X_i^2 \overset{m \to \infty}{\longrightarrow} 0.$$

If we take $\varepsilon = \frac{1}{l^2}$ then for large enough $m(l)$ and for any $n \ge m(l)$,

$$\mathbb{P}\left(|S_n - S_{m(l)}| \ge \frac{1}{l^2} \right) \le \frac{1}{l^2}. \tag{6.0.2}$$

W.l.o.g. we can assume that $m(l+1) \ge m(l)$ so that

$$\mathbb{P}\left(|S_{m(l+1)} - S_{m(l)}| \ge \frac{1}{l^2} \right) \le \frac{1}{l^2}.$$

Then,

$$\sum_{l \ge 1} \mathbb{P}\left(|S_{m(l+1)} - S_{m(l)}| \ge \frac{1}{l^2} \right) \le \sum_{l \ge 1} \frac{1}{l^2} < \infty$$

and by Borel-Cantelli lemma

$$\mathbb{P}\left(|S_{m(l+1)} - S_{m(l)}| \ge \frac{1}{l^2} \text{ i.o} \right) = 0.$$

As a result, for large enough (random) l and for $k > l$,

$$|S_{m(k)} - S_{m(l)}| \le \sum_{i \ge l} \frac{1}{i^2} < \frac{1}{l-1}.$$

This means that $(S_{m(l)})_{l \ge 1}$ is a Cauchy sequence and there exists the limit $S = \lim S_{m(l)}$. (6.0.2) implies that $S_n \to S$ in probability.

□

Example. Consider random series $\sum_{i\geq 1} \frac{\varepsilon_i}{i^\alpha}$ where $\mathbb{P}(\varepsilon_i = \pm 1) = \frac{1}{2}$. We have

$$\sum_{i\geq 1} \mathbb{E}\left(\frac{\varepsilon_i}{i^\alpha}\right)^2 = \sum_{i\geq 1} \frac{1}{i^{2\alpha}} < \infty \text{ if } \alpha > \frac{1}{2},$$

so the series converges a.s. for such α.

□

Section 7

Stopping times, Wald's identity. Another proof of SLLN.

Consider a sequence $(X_i)_{i\geq 1}$ of independent r.v.s and an integer valued random variable $V \in \{1, 2, \ldots\}$. We say that V is *independent of the future* if $\{V \leq n\}$ is independent of $\sigma((X_i)_{i\geq n+1})$. We say that V is a *stopping time* (Markov time) if $\{V \leq n\} \in \sigma(X_1, \ldots, X_n)$ for all n. Clearly, a stopping time is independent of the future. An example of stopping time is $V = \min\{k \geq 1, S_k \geq 1\}$.

Suppose that V is independent of the future. We can write

$$\begin{aligned} \mathbb{E}S_V &= \sum_{k\geq 1} \mathbb{E}S_V \mathrm{I}(V=k) = \sum_{k\geq 1} \mathbb{E}S_k \mathrm{I}(V=k) \\ &= \sum_{k\geq 1}\sum_{n\leq k} \mathbb{E}X_n \mathrm{I}(V=k) \stackrel{(*)}{=} \sum_{n\geq 1}\sum_{k\geq n} \mathbb{E}X_n \mathrm{I}(V=k) = \sum_{n\geq 1} \mathbb{E}X_n \mathrm{I}(V\geq n). \end{aligned}$$

In (*) we can interchange the order of summation if, for example, the double sequence is absolutely summable, by Fubini-Tonelli theorem. Since V is independent of the future, the event $\{V \geq n\} = \{V \leq n-1\}^c$ is independent of $\sigma(X_n)$ and we get

$$\mathbb{E}S_V = \sum_{n\geq 1} \mathbb{E}X_n \times \mathbb{P}(V \geq n). \tag{7.0.1}$$

This implies the following.

Theorem 14 *(Wald's identity.) If $(X_i)_{i\geq 1}$ are i.i.d., $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}V < \infty$, then $\mathbb{E}S_V = \mathbb{E}X_1\mathbb{E}V$.*

Proof. By (7.0.1) we have,

$$\mathbb{E}S_V = \sum_{n\geq 1} \mathbb{E}X_n\, \mathbb{P}(V \geq n) = \mathbb{E}X_1 \sum_{n\geq 1} \mathbb{P}(V \geq n) = \mathbb{E}X_1\, \mathbb{E}V.$$

The reason we can interchange the order of summation in (*) is because under our assumptions the double sequence is absolutely summable since

$$\sum_{n\geq 1}\sum_{k\geq n} \mathbb{E}|X_n| \mathrm{I}(V=k) = \sum_{n\geq 1} \mathbb{E}|X_n| \mathrm{I}(V\geq n) = \mathbb{E}|X_1|\, \mathbb{E}V < \infty,$$

so we can apply Fubini-Tonelli theorem.

□

Theorem 15 *(Markov property) Suppose that $(X_i)_{i\geq 1}$ are i.i.d. and V is a stopping time. Then $(V, X_1, \ldots, X_V)$ is independent of $(X_{V+1}, X_{V+2}, \ldots)$ and*

$$(X_{V+1}, X_{V+2}, \ldots) \stackrel{\text{dist}}{=} (X_1, X_2, \ldots),$$

where $\stackrel{\text{dist}}{=}$ *means equality in distribution.*

Proof. Given a subset $N \subseteq \mathbb{N}$ and sequences (B_i) and (C_i) of Borel sets on $\mathbb{R}$, define events

$$A = \Big\{ V \in N, X_1 \in B_1, ..., X_V \in B_V \Big\}$$

and for any $k \geq 1$,

$$D = \Big\{ X_{V+1} \in C_1, ..., X_{V+k} \in C_k \Big\}.$$

We have,

$$\mathbb{P}(DA) = \sum_{n \geq 1} \mathbb{P}\left(DA\{V = n\}\right) = \sum_{n \geq 1} \mathbb{P}\left(D_n A\{V = n\}\right)$$

where

$$D_n = \{X_{n+1} \in C_1, \ldots, X_{n+k} \in C_k\}.$$

The intersection of events

$$A\{V = n\} = \begin{cases} \emptyset, & n \notin N \\ \{V = n, X_1 \in B_1, \ldots, X_n \in B_n\}, & \text{otherwise.} \end{cases}$$

Since V is a stopping time, $\{V = n\} \in \sigma(X_1, \ldots, X_n)$ and $A\{V = n\} \in \sigma(X_1, \ldots, X_n)$. On the other hand, $D_n \in \sigma(X_{n+1}, \ldots)$ and, as a result,

$$\mathbb{P}(DA) = \sum_{n \geq 1} \mathbb{P}(D_n)\mathbb{P}(A\{V = n\}) = \sum_{n \geq 1} \mathbb{P}(D_0)\mathbb{P}(A\{V = n\}) = \mathbb{P}(D_0)\mathbb{P}(A),$$

and this finishes the proof.

□

Remark. One could be a little bit more careful when talking about the events generated by a vector $(V, X_1, \ldots, X_V)$ that has random length. In the proof we implicitly assumed that such events are generated by events

$$A = \Big\{ V \in N, X_1 \in B_1, ..., X_V \in B_V \Big\}$$

which is a rather intuitive definition. However, one could be more formal and define a σ-algebra of events generated by $(V, X_1, \ldots, X_V)$ as events A such that $A \cap \{V \leq n\} \in \sigma(X_1, \ldots, X_n)$ for any $n \geq 1$. This means that when $V \leq n$ the event A is expressed only in terms of $X_1, \ldots, X_n$. It is easy to check that with this more formal definition the proof remains exactly the same.

□

Let us give one interesting application of Markov property and Wald's identity that will yield another proof of strong law of large numbers.

Theorem 16 *Suppose that* $(X_i)_{i \geq 1}$ *are i.i.d. such that* $\mathbb{E}X_1 > 0$. *If* $Z = \inf_{n \geq 1} S_n$ *then* $\mathbb{P}(Z > -\infty) = 1$. *(Partial sums can not drift down to* $-\infty$ *if* $\mathbb{E}X_1 > 0$. *Of course, this is obvious by SLLN.)*

Proof. Let us define (see figure 7.1),

$$\tau_1 = \min\{k \geq 1, S_k \geq 1\},\ Z_1 = \min_{k \leq \tau_1} S_k,\ S_k^{(2)} = S_{\tau_1 + k} - S_{\tau_1},$$

$$\tau_2 = \min\left\{k \geq 1, S_k^{(2)} \geq 1\right\},\ Z_2 = \min_{k \leq \tau_2} S_k^{(2)},\ S_k^{(3)} = S_{\tau_2 + k}^{(2)} - S_{\tau_2}^{(2)}.$$

By induction,

$$\tau_n = \min\left\{k \geq 1, S_k^{(n)} \geq 1\right\},\ Z_n = \min_{k \leq \tau_n} S_k^{(n)},\ S_k^{(n+1)} = S_{\tau_n + k}^{(n)} - S_{\tau_n}^{(n)}.$$

$Z_1, ..., Z_n$ are i.i.d. by Markov property.

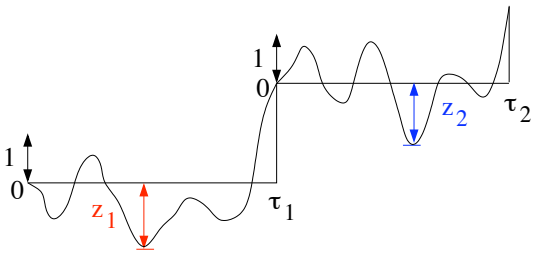


Figure 7.1: A sequence of stopping times.

Notice that, by construction, $S_{\tau_1+\cdots+\tau_{n-1}} \geq n-1$ and

$$Z = \inf_{k\geq 1} S_k = \inf\{Z_1, S_{\tau_1} + Z_2, S_{\tau_1+\tau_2} + Z_3, ...\}.$$

We have,

$$\{Z \leq -N\} = \bigcup_{k\geq 1}\{S_{\tau_1+\ldots+\tau_{k-1}} + Z_k \leq -N\} \subseteq \bigcup_{k\geq 1}\{k-1+Z_k \leq -N\}.$$

Therefore,

$$\mathbb{P}(Z \leq -N) \leq \sum_{k\geq 1}\mathbb{P}(k-1+Z_k \leq -N) = \sum_{k\geq 1}\mathbb{P}(Z_k \leq -N-k+1)$$

$$= \sum_{k\geq 1}\mathbb{P}(Z_1 \leq -N-k+1) = \sum_{j\geq N}\mathbb{P}(Z_1 \leq -j) \leq \sum_{j\geq N}\mathbb{P}(|Z_1| \geq j) \overset{N\to\infty}{\longrightarrow} 0$$

if we can show that $\mathbb{E}|Z_1| < \infty$ since

$$\sum_{j\geq 1}\mathbb{P}(|Z_1| \geq j) \leq \mathbb{E}|Z_1| < \infty.$$

We can write

$$\mathbb{E}|Z_1| \leq \mathbb{E}\sum_{i\leq\tau_1}|X_i| \overset{\text{Wald}}{=} \mathbb{E}|X_1|\mathbb{E}\tau_1 < \infty$$

if we can show that $\mathbb{E}\tau_1 < \infty$. This is left as an exercise (hint: truncate X_i's and τ_1 and use Wald's identity). We proved that $\mathbb{P}(Z \leq -N) \overset{N\to\infty}{\longrightarrow} 0$ which, of course, implies that $\mathbb{P}(Z > -\infty) = 1$.

□

This result gives another proof of the SLLN.

Theorem 17 *If $(X_i)_{i\geq 1}$ are i.i.d. and $\mathbb{E}X_1 = 0$ then $\frac{S_n}{n} \to 0$ a.s.*

Proof. Given $\varepsilon > 0$ we define $X_i^\varepsilon = X_i + \varepsilon$ so that $\mathbb{E}X_1^\varepsilon = \varepsilon > 0$. By the above result, $\inf_{n\geq 1}(S_n + n\varepsilon) > -\infty$ with probability one. This means that for all $n \geq 1$, $S_n + n\varepsilon \geq -M > -\infty$ for some large enough M. Dividing both sides by n and letting $n \to \infty$ we get

$$\liminf_{n\to\infty}\frac{S_n}{n} \geq -\varepsilon$$

with probability one. We can then let $\varepsilon \to 0$ over some sequence. Similarly, we prove that $\limsup \frac{S_k}{k} \leq 0$ with probability one.

□

Section 8

Convergence of Laws. Selection Theorem.

In this section we will begin the discussion of weak convergence of distributions on metric spaces. Let (S, d) be a metric space with a metric d. Consider a measurable space $(S, \mathcal{B})$ with Borel σ-algebra $\mathcal{B}$ generated by open sets and let $(\mathbb{P}_n)_{n\geq 1}$ and $\mathbb{P}$ be probability distributions on $\mathcal{B}$. We define

$$C_b(S) = \{f : S \to \mathbb{R} \text{ - continuous and bounded}\}.$$

We say that $\mathbb{P}_n \to \mathbb{P}$ *weakly* if

$$\int f d\mathbb{P}_n \to \int f d\mathbb{P} \quad \text{for all} \quad f \in C_b(S).$$

Theorem 18 *If $S = \mathbb{R}$ then $\mathbb{P}_n \to \mathbb{P}$ iff*

$$F_n(t) = \mathbb{P}_n\big((-\infty, t]\big) \to F(t) = \mathbb{P}\big((-\infty, t]\big)$$

for any point of continuity t of $F(t)$.

Proof. "$\Longrightarrow$" Let us approximate an indicator function by a continuous functions as in figure 8.1, i.e.

$$\varphi_1(X) \leq \mathrm{I}(X \leq t) \leq \varphi_2(X), \;\; \varphi_1, \varphi_2 \in C_b(\mathbb{R}).$$

For convenience of notations, instead of writing integrals w.r.t. $\mathbb{P}_n$ we will write expectations of a r.v. X_n

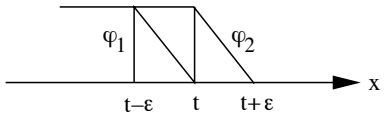


Figure 8.1: Approximating indicator.

with distribution $\mathbb{P}_n$.

$$\mathbb{P}(X \leq t-\varepsilon) \leq \mathbb{E}\varphi_1(X) \leftarrow \mathbb{E}\varphi_1(X_n) \leq F_n(t) = \mathbb{P}(X_n \leq t) \leq \mathbb{E}\varphi_2(X_n) \to \mathbb{E}\varphi_2(X) \leq \mathbb{P}(X \leq t+\varepsilon)$$

as $n \to \infty$. Therefore, for any $\varepsilon > 0$,

$$F(t-\varepsilon) \leq \liminf F_n(t) \leq \limsup F_n(t) \leq F(t+\varepsilon).$$

Since t is a point of continuity of F, letting $\varepsilon \to 0$ proves the result.

"$\Longleftarrow$" Let $PC(F)$ be the set of points of continuity of F. Since F is monotone, the set $PC(F)$ is dense in $\mathbb{R}$. Take M large enough such that both $M, -M \in PC(F)$ and $\mathbb{P}([-M,M]^c) \le \varepsilon$. Clearly, for large enough k we have $\mathbb{P}_k([-M,M]^c) \le 2\varepsilon$. For any $n > 1$, take a sequence of points

$$-M = x_1^n \le x_2^n \le \cdots \le x_n^n = M$$

such that all $x_i \in PC(F)$ and $\max_i |x_{i+1}^n - x_i^n| \to 0$ as $n \to \infty$. Given a function $f \in C_b(\mathbb{R})$, consider an approximating function

$$f_n(x) = \sum_{1<i\le n} f(x_i)\mathrm{I}(x \in (x_{i-1}^n, x_i^n]) + 0 \cdot \mathrm{I}(x \notin [-M,M]).$$

Since f in continuous,

$$\sup_{|x|\le M} |f_n(x) - f(x)| \le \delta_n \to 0, \;\; n \to \infty.$$

Since all x_i^n are in $PC(F)$ we can write

$$\mathbb{E}f_n(X_k) = \sum_{1<i\le n} f_n(x_i^n)\Big(F_k(x_i^n) - F_k(x_{i-1}^n\Big) \stackrel{k\to\infty}{\longrightarrow} \sum_{1<i\le n} f_n(x_i^n)\Big(F(x_i^n) - F(x_{i-1}^n)\Big) = \mathbb{E}f_n(X).$$

Also,

$$\Big|\mathbb{E}f(X) - \mathbb{E}f_n(X)\Big| \le \|f\|_\infty \mathbb{P}(X \notin [-M,M]) + \delta_n \le \|f\|_\infty \varepsilon + \delta_n$$

and, similarly,

$$\Big|\mathbb{E}f(X_k) - \mathbb{E}f_n(X_k)\Big| \le \|f\|_\infty \mathbb{P}(X_k \notin [-M,M]) + \delta_n \le \|f\|_\infty 2\varepsilon + \delta_n.$$

Letting $k \to \infty$, $\varepsilon \to 0$ and $n \to \infty$ proves that $\mathbb{E}f(X_k) \to \mathbb{E}f(X)$.

□

We say that a sequence of distributions $(\mathbb{P}_n)_{n\ge1}$ on a metric space (S,d) is *uniformly tight* if for any $\varepsilon > 0$ there exists a compact $K \subseteq S$ such that $\mathbb{P}_n(K) > 1 - \varepsilon$ for all n. Our next goal is to prove the following theorem.

Theorem 19 *(Selection theorem) If $(\mathbb{P}_n)_{n\ge1}$ is a uniformly tight sequence of laws on the metric space (S,d) then there exists a subsequence $(n(k))$ such that $\mathbb{P}_{n(k)}$ converges weakly to some probability law $\mathbb{P}$.*

We will prove the Selection Theorem for arbitrary metric spaces, since this result will be useful to us later when we study the convergence of laws on general metric spaces. However, when $S = \mathbb{R}^k$ one can proceed in a more intuitive way, based on the following Lemma.

Lemma 12 *(Cantor's diagonalization) Let A be a countable $\subseteq S$ and $f_n : S \to \mathbb{R}, n \ge 1$. Then there exists a subsequence $(n(k))$ such that $f_{n(k)}(a)$ converges for all $a \in A$ (if $f_{n(k)}(a)$ is unbounded, maybe to $\pm\infty$).*

Proof. Let $A = \{a_1, a_2, \ldots\}$. Take $(n^1(k))$ such that $f_{n^1(k)}(a_1)$ converges. Take $(n^2(k)) \subseteq (n^1(k))$ such that $f_{n^2(k)}(a_2)$ converges. By induction, take $(n^l(k)) \subseteq (n^{l-1}(k))$ such that $f_{n^l(k)}(a_l)$ converges. Now consider a sequence $(n^k(k))$. Clearly, $f_{n^k(k)}(a_l)$ converges for any l because for $k \ge l, n^k(k) \in \{n^l(k)\}$ by construction.

□

Define a joint c.d.f. on $\mathbb{R}^k$ by

$$F(t) = \mathbb{P}(X_1 \le t_1, \ldots, X_k \le t_k) \;\text{ where }\; t = (t_1, \ldots, t_k).$$

Let A be a dense set of points in $\mathbb{R}^k$. By Lemma, there exists a subsequence $(n(k))$ such that $F_{n(k)}(a) \to F(a)$ for all $a \in A$. For $x \in \mathbb{R}^k \setminus A$ we can extend F by

$$F(x) = \inf\{F(a) : a \in A, x_i < a_i\}.$$

$F(x)$ is a c.d.f. on $\mathbb{R}^k$ (exercise). The fact that $\mathbb{P}_n$ are uniformly tight ensures that $F(x) \to 0$ or 1 if all $x_i \to -\infty$ or $+\infty$. Let x be a point of continuity of $F(x)$ and let $a, b \in A$ such that $a_i < x_i < b_i$ for all i. We have,

$$F(a) \leftarrow F_{n(k)}(a) \leq F_{n(k)}(x) \leq F_{n(k)}(b) \to F(b)$$

as $k \to \infty$. Since x is a point of continuity and A is dense,

$$F(a) \stackrel{a \to x}{\longrightarrow} F(x), \;\; F(b) \stackrel{b \to x}{\longrightarrow} F(x),$$

and this proves that $F_{n(k)}(x) \to F(x)$ for all such x. Similarly to one-dimensional case one can show that for any $f \in C_b(\mathbb{R}^k)$,

$$\int f dF_{n(k)} \to \int f dF.$$

Proof of Theorem 19. If K is a compact then $C_b(K) = C(K)$. Later in these lectures, when we deal in more detail with convergence on general metric spaces, we will prove the following fact which is well-known and is a consequence of the Stone-Weierstrass theorem.

Fact. $C(K)$ is separable w.r.t. ℓ_∞ norm $||f||_\infty = \sup_{x \in K} |f(x|$.

Even though we are proving Selection theorem for a general metric space, right now we are mostly interested in the case $S = \mathbb{R}^k$ where this fact is a simple consequence of the Weierstrass theorem that any continuous function can be approximated by polynomials.

Since $\mathbb{P}_n$ are uniformly tight, for any $r \geq 1$ we can find a compact K_r such that $\mathbb{P}_n(K_r) > 1 - \frac{1}{r}$. Let $C_r \subset C(K_r)$ be a countable and dense subset of $C(K_r)$. By Cantor's diagonalization argument there exists a subsequence $(n(k))$ such that $\mathbb{P}_{n(k)}(f)$ converges for all $f \in C_r$ for all $r \geq 1$. Since C_r is dense in $C(K_r)$ this implies that $\mathbb{P}_{n(k)}(f)$ converges for all $f \in C(K_r)$ for all $r \geq 1$. Next, for any $f \in C_b(S)$,

$$\left| \int f d\mathbb{P}_{n(k)} - \int_{K_r} f d\mathbb{P}_{n(k)} \right| \leq \int_{K_r^c} |f| d\mathbb{P}_{n(k)} \leq ||f||_\infty \mathbb{P}_{n(k)}(K_r^c) \leq \frac{||f||_\infty}{r}.$$

This implies that the limit

$$I(f) := \lim_{k \to \infty} \int f d\mathbb{P}_{n(k)} \tag{8.0.1}$$

exists. The question is why this limit is an integral over some probability measure $\mathbb{P}$? On each of the compacts K_r we could use Riesz's representation theorem for continuous functionals on $C(K_r)$ and then extend this representation to the union of K_r. Instead, we will prove this as a consequence of a more general result, the Stone-Daniell theorem from measure theory, which says the following.

A family of function $\alpha = \{f : S \to \mathbb{R}\}$ is called a *vector lattice* if

$$f, g \in \alpha \Longrightarrow cf + g \in \alpha, \forall c \in \mathbb{R} \quad \text{and} \quad f \wedge g, f \vee g \in \alpha.$$

A functional $I : \alpha \to \mathbb{R}$ is called a *pre-integral* if

1. $I(cf + g) = cI(f) + I(g)$,
2. $f \geq 0, I(f) \geq 0$,
3. $f_n \downarrow 0, ||f_n||_\infty < \infty \Longrightarrow I(f_n) \to 0$.

Theorem 20 *(Stone-Daniell) If α is a vector lattice and I is a pre-integral on α then $I(f) = \int f d\mu$ for some unique measure μ on σ-algebra generated by functions in α (i.e. minimal σ-algebra on which all functions in α are measurable).*

We will use this theorem with $\alpha = C_b(S)$ and I defined in (8.0.1). The first two properties are obvious. To prove the third one let us consider a sequence such that

$$f_n \downarrow 0, \;\; 0 \leq f_n(x) \leq f_1(x) \leq ||f_1||_\infty.$$

On any compact K_r, $f_n \downarrow 0$ uniformly, i.e.

$$||f_n||_{\infty,K_r} \le \varepsilon_{n,r} \overset{n\to\infty}{\longrightarrow} 0.$$

Since

$$\int f_n d\mathbb{P}_{n(k)} = \int_{K_r \cup K_r^c} f_n d\mathbb{P}_{n(k)} \le \varepsilon_{n,r} + \frac{1}{r}\|f_1||_\infty,$$

we get

$$I(f_n) = \lim_{k\to\infty} \int f_n d\mathbb{P}_{n(k)} \le \varepsilon_{n,r} + \frac{1}{r}||f_1||_\infty.$$

Letting $n \to \infty$ and $r \to \infty$ we get that $I(f_n) \to 0$. By the Stone-Daniell theorem

$$I(f) = \int f d\mathbb{P}$$

for some measure on $\sigma(C_b(S))$. The choice of $f = 1$ gives $I(f) = 1 = \mathbb{P}(S)$ which means that $\mathbb{P}$ is a probability measure. Finally, let us show that $\sigma(C_b(S)) = \mathcal{B}$ - Borel σ-algebra generated by open sets. Since any $f \in C_b(S)$ is measurable on $\mathcal{B}$ we get $\sigma(C_b(S)) \subseteq \mathcal{B}$. On the other hand, let $F \subseteq S$ be any closed set and take a function $f(x) = \min(1, d(x, F))$. We have, $|f(x) - f(y)| \le d(x, y)$ so $f \in C_b(S)$ and

$$f^{-1}(\{0\}) \in \sigma(C_b(S)).$$

However, since F is closed, $f^{-1}(\{0\}) = \{x : d(x, F) = 0\} = F$ and this proves that $\mathcal{B} \subseteq \sigma(C_b(S))$.

□

Theorem 21 *If $\mathbb{P}_n$ converges weakly to $\mathbb{P}$ on $\mathbb{R}^k$ then $(\mathbb{P}_n)_{n\ge1}$ is uniformly tight.*

Proof. For any $\varepsilon > 0$ there exists large enough $M > 0$, such that $\mathbb{P}(|x| > M) < \varepsilon$. Consider a function

$$\alpha(s) = \begin{cases} 0, & s \le M, \\ 1, & s \ge 2M, \\ \frac{1}{M}(s - M), & M \le s \le 2M. \end{cases}$$

and let $\alpha(x) := \alpha(|x|)$ for $x \in \mathbb{R}^k$. Since $\mathbb{P}_n \to \mathbb{P}$ weakly,

$$\int \alpha(x) d\mathbb{P}_n \to \int \alpha(x) d\mathbb{P}.$$

This implies that

$$\mathbb{P}_n\Big(|x| > 2M\Big) \le \int \alpha(x) d\mathbb{P}_n \to \int \alpha(x) d\mathbb{P} \le \mathbb{P}\Big(|x| > M\Big) \le \varepsilon.$$

For n large enough, $n \ge n_0$, we get $\mathbb{P}_n(|x| > 2M) \le 2\varepsilon$. For $n < n_0$ choose M_n so that $\mathbb{P}_n(|x| > M_n) \le 2\varepsilon$. Take $M' = \max\{M_1, \dots, M_{n_0-1}, 2M\}$. As a result, $\mathbb{P}_n(|x| > M') \le 2\varepsilon$ for all $n \ge 1$.

□

Lemma 13 *If for any sequence $(n(k))_{k\ge1}$ there exists a subsequence $(n(k(r)))_{r\ge1}$ such that $\mathbb{P}_{n(k(r))} \to \mathbb{P}$ weakly then $\mathbb{P}_n \to \mathbb{P}$ weakly.*

Proof. Suppose not. Then for some $f \in C_b(S)$ and for some $\varepsilon > 0$ there exists a subsequence $(n(k))$ such that

$$\Big|\int f d\mathbb{P}_{n(k)} - \int f d\mathbb{P}\Big| > \varepsilon.$$

But this contradicts the fact that for some subsequence $\mathbb{P}_{n(k(r))} \to \mathbb{P}$ weakly.

□

Consider r.v.s X and X_n on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in a metric space (S, d). Let $\mathbb{P}$ and $\mathbb{P}_n$ be their corresponding laws on Borel sets $\mathcal{B}$ in S. Convergence of X_n to X in probability and almost surely is defined exactly the same way as for $S = \mathbb{R}$ by replacing $|X_n - X|$ with $d(X_n, X)$.

Lemma 14 *$X_n \to X$ in probability iff for any sequence $(n(k))$ there exists a subsequence $(n(k(r)))$ such that $X_{n(k(r))} \to X$ a.s.*

Proof. "$\Longleftarrow$". Suppose X_n does not converge to X in probability, Then for small enough $\varepsilon > 0$ there exists a subsequence $(n(k))$ such that

$$\mathbb{P}\Big(d(X, X_{n(k)}) \geq \varepsilon\Big) \geq \varepsilon.$$

This contradicts the existence of subsequence $X_{n(k(r))}$ that converges to X a.s.

"$\Longrightarrow$". Given a subsequence $(n(k))$ let us choose $(k(r))$ so that

$$\mathbb{P}\Big(d(X_{n(k(r))}, X) \geq \frac{1}{r}\Big) \leq \frac{1}{r^2}.$$

By Borel-Cantelli lemma, these events can occur i.o. with probability 0, which means that with probability one for large enough r

$$d(X_{n(k(r))}, X) \leq \frac{1}{r},$$

i.e. $X_{n(k(r))} \to X$ a.s.

□

Lemma 15 *$X_n \to X$ in probability then $X_n \to X$ weakly.*

Proof. By Lemma 14, for any subsequence $(n(k))$ there exists a subsequence $(n(k(r)))$ such that $X_{n(k(r))} \to X$ a.s. Given $f \in C_b(\mathbb{R})$, by dominated convergence theorem,

$$\mathbb{E}f(X_{n(k(r))}) \to \mathbb{E}f(X),$$

i.e. $X_{n(k(r))} \to X$ weakly. By Lemma 13, $X_n \to X$ weakly.

□

Section 9

Characteristic Functions. Central Limit Theorem on $\mathbb{R}$.

Let $X = (X_1, \ldots, X_k)$ be a random vector on $\mathbb{R}^k$ with distribution $\mathbb{P}$ and let $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$. *Characteristic function* of X is defined by

$$f(t) = \mathbb{E}e^{i(t,X)} = \int e^{i(t,x)} d\mathbb{P}(x).$$

If X has standard normal distribution $\mathcal{N}(0,1)$ and $\lambda \in \mathbb{R}$ then

$$\mathbb{E}e^{\lambda X} = \frac{1}{\sqrt{2\pi}} \int e^{\lambda x - \frac{x^2}{2}} dx = e^{\frac{\lambda^2}{2}} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\lambda)^2}{2}} dx = e^{\frac{\lambda^2}{2}}.$$

For complex $\lambda = it$, consider analytic function

$$\varphi(x) = e^{itx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{for} \quad x \in \mathbb{C}.$$

By Cauchy's theorem, integral over a closed path is equal to 0. Let us take a closed path $x + i0$ for x from $-\infty$ to $+\infty$ and $x + it$ for x from $+\infty$ to $-\infty$. Then

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx - \frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{it(it+x) - \frac{1}{2}(it+x)^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2 + itx + \frac{1}{2}t^2 - itx - \frac{1}{2}x^2} dx = e^{-\frac{t^2}{2}} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = e^{-\frac{t^2}{2}}. \end{aligned} \tag{9.0.1}$$

If Y has normal distribution $\mathcal{N}(m, \sigma^2)$ then

$$\mathbb{E}e^{itY} = \mathbb{E}e^{it(m+\sigma X)} = e^{itm - \frac{t^2\sigma^2}{2}}.$$

Lemma 16 *If X is a real-valued r.v. such that $\mathbb{E}|X|^r < \infty$ for integer r then $f(t) \in C^r(\mathbb{R})$ and*

$$f^{(j)}(t) = \mathbb{E}(iX)^j e^{itX}$$

for $j \le r$.

Proof. If $r = 0$, then $|e^{itX}| \le 1$ implies

$$f(t) = \mathbb{E}e^{itX} \to \mathbb{E}e^{isX} = f(s) \quad \text{if} \quad t \to s,$$

by dominated convergence theorem. This means that $f \in C(\mathbb{R})$. If $r = 1$, $\mathbb{E}|X| < \infty$, we can use

$$\left| \frac{e^{itX} - e^{isX}}{t - s} \right| \leq |X|$$

and, therefore, by dominated convergence theorem,

$$f'(t) = \lim_{s \to t} \mathbb{E} \frac{e^{itX} - e^{isX}}{t - s} = \mathbb{E} i X e^{itX}.$$

Also, by dominated convergence theorem, $\mathbb{E} i X e^{itX} \in C(\mathbb{R})$, which means that $f \in C^1(\mathbb{R})$. We proceed by induction. Suppose that we proved that

$$f^{(j)}(t) = \mathbb{E}(iX)^j e^{itX}$$

and that $r = j + 1$, $\mathbb{E}|X|^{j+1} < \infty$. Then, we can use that

$$\left| \frac{(iX)^j e^{itX} - (iX)^j e^{isX}}{t - s} \right| \leq |X|^{j+1}$$

so that by dominated convergence theorem $f^{(j+1)}(t) = \mathbb{E}(iX)^{j+1} e^{itX} \in C(\mathbb{R})$.

□

The main goal of this section is to prove one of the most famous results in Probability Theory.

Theorem 22 *(Central Limit Theorem) Consider an i.i.d. sequence $(X_i)_{i \geq 1}$ such that $\mathbb{E}X_1 = 0, \mathbb{E}X_1^2 = \sigma^2 < \infty$ and let $S_n = \sum_{i \leq n} X_i$. Then $S_n/\sqrt{n}$ converges in distribution to $\mathcal{N}(0, \sigma^2)$.*

We will start with the following.

Lemma 17 *We have,*

$$\lim_{n \to \infty} \mathbb{E} e^{it \frac{S_n}{\sqrt{n}}} = e^{-\frac{1}{2}\sigma^2 t^2}.$$

Proof. By independence,

$$\mathbb{E} e^{it \frac{S_n}{\sqrt{n}}} = \prod_{i \leq n} \mathbb{E} e^{\frac{itX_i}{\sqrt{n}}} = \left(\mathbb{E} e^{\frac{itX_1}{\sqrt{n}}} \right)^n.$$

Since $\mathbb{E}X_1^2 < \infty$ previous lemma implies that $\varphi(t) \in C^2(\mathbb{R})$ and, therefore,

$$\varphi(t) = \mathbb{E} e^{itX_1} = \varphi(0) + \varphi'(0)t + \frac{1}{2}\varphi''(0)t^2 + o(t^2) \ \text{ as } \ t \to 0.$$

Since

$$\varphi(0) = 1, \ \varphi'(0) = \mathbb{E} i X e^{i \cdot 0 \cdot X} = i\mathbb{E}X = 0, \ \varphi''(0) = \mathbb{E}(iX)^2 = -\mathbb{E}X^2 = -\sigma^2$$

we get

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2).$$

Finally,

$$\mathbb{E} e^{\frac{itS_n}{\sqrt{n}}} = \left(\varphi\Big(\frac{t}{\sqrt{n}} \Big) \right)^n = \left(1 - \frac{\sigma^2 t^2}{2n} + o\Big(\frac{t^2}{n} \Big) \right)^n \to e^{-\frac{1}{2}\sigma^2 t^2}, \ \ n \to \infty.$$

□

Next, we want to show that characteristic function uniquely determines the distribution. Let $X \sim \mathbb{P}, Y \sim \mathbb{Q}$ be two independent random vectors on $\mathbb{R}^k$. We denote by $\mathbb{P} * \mathbb{Q}$ the *convolution* of $\mathbb{P}$ and $\mathbb{Q}$ which is the law $\mathcal{L}(X + Y)$ of the sum $X + Y$. We have,

$$\begin{aligned} \mathbb{P} * \mathbb{Q}(A) = \mathbb{E} I(X + Y \in A) &= \iint \mathrm{I}(x + y \in A) d\mathbb{P}(x) d\mathbb{Q}(y) \\ &= \iint \mathrm{I}(x \in A - y) d\mathbb{P}(x) d\mathbb{Q}(y) = \int \mathbb{P}(A - y) d\mathbb{Q}(y). \end{aligned}$$

If $\mathbb{P}$ has density p then

$$\begin{aligned}\mathbb{P}*\mathbb{Q}(A) &= \iint \mathrm{I}(x+y\in A)p(x)dxd\mathbb{Q}(y) = \iint \mathrm{I}(z\in A)p(z-y)dzd\mathbb{Q}(y)\\ &= \iint\limits_A p(z-y)dzd\mathbb{Q}(y) = \int\limits_A \Big(\int p(z-y)d\mathbb{Q}(y)\Big)dz\end{aligned}$$

which means that $\mathbb{P}*\mathbb{Q}$ has density

$$f(x) = \int p(x-y)d\mathbb{Q}(y). \tag{9.0.2}$$

If, in addition, $\mathbb{Q}$ has density q then

$$f(x) = \int p(x-y)q(y)dy.$$

Denote by $\mathcal{N}(0,\sigma^2 I)$ the law of the random vector $X=(X_1,\ldots,X_k)$ of i.i.d. $\mathcal{N}(0,\sigma^2)$ random variables whose density on $\mathbb{R}^k$ is

$$\prod_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}x_i^2} = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^k e^{-\frac{1}{2\sigma^2}|x|^2}.$$

For a distribution $\mathbb{P}$ denote $\mathbb{P}^\sigma = \mathbb{P}*\mathcal{N}(0,\sigma^2 I)$.

Lemma 18 *$\mathbb{P}^\sigma = \mathbb{P}*\mathcal{N}(0,\sigma^2 I)$ has density*

$$p^\sigma(x) = \Big(\frac{1}{2\pi}\Big)^k \int f(t)e^{-i(t,x)-\frac{\sigma^2}{2}|t|^2}dt$$

where $f(t)=\int e^{i(t,x)}d\mathbb{P}(x)$.

Proof. By (9.0.2), $\mathbb{P}*\mathcal{N}(0,\sigma^2 I)$ has density

$$p^\sigma(x) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^k \int e^{-\frac{1}{2\sigma^2}|x-y|^2}d\mathbb{P}(y).$$

Using (9.0.1), we can write

$$e^{-\frac{1}{2\sigma^2}(x_i-y_i)^2} = \frac{1}{\sqrt{2\pi}}\int e^{-i\frac{1}{\sigma}(x_i-y_i)z_i}e^{-\frac{1}{2}z_i^2}dz_i$$

and taking a product over $i\le k$ we get

$$e^{-\frac{1}{2\sigma^2}|x-y|^2} = \Big(\frac{1}{\sqrt{2\pi}}\Big)^k \int e^{-i\frac{1}{\sigma}(x-y,z)}e^{-\frac{1}{2}|z|^2}dz.$$

Then we can continue

$$\begin{aligned}p^\sigma(x) &= \Big(\frac{1}{2\pi\sigma}\Big)^k \iint e^{-i\frac{1}{\sigma}(x-y,z)-\frac{1}{2}|z|^2}dzd\mathbb{P}(y)\\ &= \Big(\frac{1}{2\pi\sigma}\Big)^k \iint e^{-i\frac{1}{\sigma}(x-y,z)-\frac{1}{2}|z|^2}d\mathbb{P}(y)dz\\ &= \Big(\frac{1}{2\pi\sigma}\Big)^k \int f\Big(\frac{z}{\sigma}\Big)e^{-i\frac{1}{\sigma}(x,z)-\frac{1}{2}|z|^2}dz.\end{aligned}$$

Let $z=t\sigma$.

□

Theorem 23 *(Uniqueness) If*

$$\int e^{i(t,x)} d\mathbb{P}(x) = \int e^{i(t,x)} d\mathbb{Q}(x)$$

then $\mathbb{P} = \mathbb{Q}$.

Proof. By the above Lemma, $\mathbb{P}^\sigma = \mathbb{Q}^\sigma$. If $X \sim \mathbb{P}$ and $\mu \sim N(0, I)$ then $X + \sigma\mu \to X$ almost surely as $\sigma \to 0$ and, therefore, $\mathbb{P}^\sigma \to \mathbb{P}$ weakly. Similarly, $\mathbb{Q}^\sigma \to \mathbb{Q}$.

□

We proved that the characteristic function of $S_n/\sqrt{n}$ converges to the c.f. of $\mathcal{N}(0, \sigma^2)$. Also, the sequence

$$\left(\mathcal{L}\Big(\frac{S_n}{\sqrt{n}} \Big) \right)_{n \geq 1} \text{ - is uniformly tight,}$$

since by Chebyshev's inequality

$$\mathbb{P}\Big(\Big| \frac{S_n}{\sqrt{n}} \Big| > M \Big) \leq \frac{\sigma^2}{M^2} < \varepsilon$$

for large enough M. To finish the proof of the CLT on the real line we apply the following.

Lemma 19 *If* $(\mathbb{P}_n)$ *is uniformly tight and*

$$f_n(t) = \int e^{itx} d\mathbb{P}_n(x) \to f(t)$$

then $\mathbb{P}_n \to \mathbb{P}$ *and* $f(t) = \int e^{itx} d\mathbb{P}(x)$.

Proof. For any sequence $(n(k))$, by Selection Theorem, there exists a subsequence $(n(k(r)))$ such that $\mathbb{P}_{n(k(r))}$ converges weakly to some distribution $\mathbb{P}$. Since $e^{i(t,x)}$ is bounded and continuous,

$$\int e^{i(t,x)} d\mathbb{P}_{n(k(r))} \to \int e^{i(t,x)} d\mathbb{P}(x)$$

as $r \to \infty$ and, therefore, f is a c.f. of $\mathbb{P}$. By uniqueness theorem, distribution $\mathbb{P}$ does not depend on the sequence $(n(k))$. By Lemma 13, $\mathbb{P}_n \to \mathbb{P}$ weakly.

□

Section 10

Multivariate normal distributions and CLT.

Let $\mathbb{P}$ be a probability distribution on $\mathbb{R}^k$ and let

$$g(t) = \int e^{i(t,x)} d\mathbb{P}(x).$$

We proved that $\mathbb{P}^\sigma = \mathbb{P} * \mathcal{N}(0, \sigma^2 I)$ has density

$$p^\sigma(x) = (2\pi)^{-k} \int g(t) e^{-i(t,x) - \frac{1}{2}\sigma^2 |t|^2} dt.$$

Lemma 20 *(Fourier inversion formula) If $\int |g(t)| dt < \infty$ then $\mathbb{P}$ has density*

$$p(x) = (2\pi)^{-k} \int g(t) e^{-i(t,x)} dt.$$

Proof. Since

$$g(t) e^{-i(t,x) - \frac{1}{2}\sigma^2 |t|^2} \to g(t) e^{-i(t,x)}$$

pointwise as $\sigma \to 0$ and

$$\left| g(t) e^{-i(t,x) - \frac{1}{2}\sigma^2 |t|^2} \right| \leq |g(t)| \text{ - integrable,}$$

by dominated convergence theorem $p^\sigma(x) \to p(x)$. Since $\mathbb{P}^\sigma \to \mathbb{P}$ weakly, for any $f \in C_b(\mathbb{R}^k)$,

$$\int f(x) p^\sigma(x) dx \to \int f(x) d\mathbb{P}(x).$$

On the other hand, since

$$|p^\sigma(x)| \leq (2\pi)^{-k} \int |g(t)| dt < \infty,$$

by dominated convergence theorem, for any compactly supported $f \in C_b(\mathbb{R}^k)$,

$$\int f(x) p^\sigma(x) dx \to \int f(x) p(x) dx.$$

Therefore, for any such f,

$$\int f(x) d\mathbb{P}(x) = \int f(x) p(x) dx.$$

It is now a simple exercise to show that for any bounded open set U,

$$\int_U d\mathbb{P}(x) = \int_U p(x)dx.$$

This means that $\mathbb{P}$ restricted to bounded sets has density $p(x)$ and, hence, on entire $\mathbb{R}^k$.

□

For a random vector $X = (X_1, \dots, X_k) \in \mathbb{R}^k$ we denote $\mathbb{E}X = (\mathbb{E}X_1, \dots, \mathbb{E}X_k)$.

Theorem 24 *Consider a sequence $(X_i)_{i\geq 1}$ of i.i.d. random vectors on $\mathbb{R}^k$ such that $\mathbb{E}X_1 = 0, \mathbb{E}|X_1|^2 < \infty$. Then $\mathcal{L}\Big(\frac{S_n}{\sqrt{n}}\Big)$ converges weakly to distribution $\mathbb{P}$ which has characteristic function*

$$f_p(t) = e^{-\frac{1}{2}(Ct,t)} \quad \textit{where} \quad C = \mathrm{Cov}(X_1) = \Big(\mathbb{E}X_{1,i}X_{1,j}\Big)_{1\leq i,j\leq k}. \tag{10.0.1}$$

Proof. Consider any $t \in \mathbb{R}^k$. Then $Z_i = (t, X_i)$ is i.i.d. real-valued sequence and by the proof of the CLT on the real line,

$$\mathbb{E}e^{i\left(t, \frac{S_n}{\sqrt{n}}\right)} = \mathbb{E}e^{i\frac{1}{\sqrt{n}}\sum_i (t,X_i)} \to e^{-\frac{1}{2}\mathrm{Var}((t,X_i))} = e^{-\frac{1}{2}(Ct,t)}$$

as $n \to \infty$, since

$$\mathrm{Var}\Big(t_1X_{1,1} + \cdots + t_kX_{1,k}\Big) = \sum_{i,j} t_it_j\mathbb{E}X_{1,i}X_{1,j} = (Ct,t) = t^TCt.$$

The sequence $\Big\{\mathcal{L}\Big(\frac{S_n}{\sqrt{n}}\Big)\Big\}$ is uniformly tight on $\mathbb{R}^k$ since

$$\mathbb{P}\Big(\Big|\frac{S_n}{\sqrt{n}}\Big| \geq M\Big) \leq \frac{1}{M^2}\mathbb{E}\Big|\frac{S_n}{\sqrt{n}}\Big|^2 = \frac{1}{nM^2}\mathbb{E}\left|(S_{n,1}, \dots, S_{n,k})\right|^2 = \frac{1}{nM^2}\sum_{i\leq k}\mathbb{E}S_{n,i}^2 = \frac{1}{M^2}\mathbb{E}|X_1|^2 \overset{M\to\infty}{\longrightarrow} 0.$$

It remains to apply Lemma 19 from previous section.

□

The covariance matrix $C = \mathrm{Cov}(X)$ is symmetric and non-negative definite, $(Ct,t) = \mathbb{E}(t,X)^2 \geq 0$.

The unique distribution with c.f. in (10.0.1) is called a multivariate normal distribution with covariance C and is denoted by $\mathcal{N}(0,C)$. It can also be defined more constructively as follows. Consider an i.i.d. sequence $g_1, \dots, g_n$ of $\mathcal{N}(0,1)$ r.v. and let $g = (g_1, \dots, g_n)^T$. Given a $k \times n$ matrix the covariance matrix of $Ag \in \mathbb{R}^k$ is

$$C = \mathrm{Cov}(Ag) = \mathbb{E}Ag(Ag)^T = A\mathbb{E}gg^TA^T = AA^T.$$

The c.f. of Ag is

$$\mathbb{E}e^{i(t,Ag)} = \mathbb{E}e^{i(A^Tt,g)} = e^{-\frac{1}{2}|A^Tt|^2} = e^{-\frac{1}{2}t^TAA^Tt} = e^{-\frac{1}{2}(Ct,t)}.$$

This means that $Ag \sim \mathcal{N}(0,C)$. Interestingly, the distribution of Ag depends only on AA^T and does not depend on the choice of n and A.

Exercise. Show constructively, using linear algebra, that the distribution of Ag and Bg' is the same if $AA^T = BB^T$.

□

On the other hand, given a covariance matrix C one can always find A such that $C = AA^T$. For example, let $C = QDQ^T$ be its eigenvalue decomposition for orthogonal matrix Q and diagonal D. Since C is nonnegative definite, the elements of D are nonnegative. Then, one can take $n = k$ and $A = C^{1/2} := QD^{1/2}Q^T$ or $A = QD^{1/2}$.

Density in the invertible case. Suppose $\det(C) \neq 0$. Take A such that $C = AA^T$ so that $Ag \sim \mathcal{N}(0,C)$. Since the density of g is

$$\prod_{i\leq k}\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}x_i^2\Big) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^k\exp\Big(-\frac{1}{2}|x|^2\Big),$$

for any set $\Omega \subseteq \mathbb{R}^k$ we can write

$$\mathbb{P}(Ag \in \Omega) = \mathbb{P}(g \in A^{-1}\Omega) = \int_{A^{-1}\Omega} \Big(\frac{1}{\sqrt{2\pi}}\Big)^k \exp\Big(-\frac{1}{2}|x|^2\Big)dx.$$

Let us now make the change of variables $y = Ax$ or $x = A^{-1}y$. Then

$$\mathbb{P}(Ag \in \Omega) = \int_{\Omega} \Big(\frac{1}{\sqrt{2\pi}}\Big)^k \exp\Big(-\frac{1}{2}|A^{-1}y|^2\Big) \frac{1}{|\det(A)|}dy.$$

But since

$$\det(C) = \det(AA^T) = \det(A)\det(A^T) = \det(A)^2$$

we have $|\det(A)| = \sqrt{\det(C)}$. Also

$$|A^{-1}y|^2 = (A^{-1}y)^T(A^{-1}y) = y^T(A^T)^{-1}A^{-1}y = y^T(AA^T)^{-1}y = y^TC^{-1}y.$$

Therefore, we get

$$\mathbb{P}(Ag \in \Omega) = \int_{\Omega} \Big(\frac{1}{\sqrt{2\pi}}\Big)^k \frac{1}{\sqrt{\det(C)}} \exp\Big(-\frac{1}{2}y^TC^{-1}y\Big)dy.$$

This means that the distribution $\mathcal{N}(0, C)$ has the density

$$\Big(\frac{1}{\sqrt{2\pi}}\Big)^k \frac{1}{\sqrt{\det(C)}} \exp\Big(-\frac{1}{2}y^TC^{-1}y\Big).$$

□

General case. Let us take, for example, a vector $X = QD^{1/2}g$ for i.i.d. standard normal vector g so that $X \sim \mathcal{N}(0, C)$. If $q_1, \ldots, q_k$ are the column vectors of Q then

$$X = QD^{1/2}g = (\lambda_1^{1/2}g_1)q_1 + \ldots + (\lambda_k^{1/2}g_n)q_k.$$

Therefore, in the orthonormal coordinate basis $q_1, \ldots, q_k$ a random vector X has coordinates $\lambda_1^{1/2}g_1, \ldots, \lambda_k^{1/2}g_k$. These coordinates are independent with normal distributions with variances $\lambda_1, \ldots, \lambda_k$ correspondingly. When $\det(C) = 0$, i.e. C is not invertible, some of its eigenvalues will be zero, say, $\lambda_{n+1} = \ldots = \lambda_k = 0$. Then the random X vector will be concentrated on the subspace spanned by vectors $q_1, \ldots, q_n$ but it will not have density on the entire space $\mathbb{R}^k$. On the subspace spanned by vectors $q_1, \ldots, q_n$ a vector X will have a density

$$f(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\lambda_i}} \exp\Big(-\frac{x_i^2}{2\lambda_i}\Big).$$

□

Let us look at a couple of properties of normal distributions.

Lemma 21 *If $X \sim \mathcal{N}(0, C)$ on $\mathbb{R}^k$ and $A : \mathbb{R}^k \to \mathbb{R}^m$ is linear then $AX \sim \mathcal{N}(0, ACA^T)$ on $\mathbb{R}^m$.*

Proof. The c.f. of AX is

$$\mathbb{E}e^{i(t,AX)} = \mathbb{E}e^{i(A^Tt,X)} = e^{-\frac{1}{2}(CA^Tt,A^Tt)} = e^{-\frac{1}{2}(ACA^Tt,t)}.$$

□

Lemma 22 *X is normal on $\mathbb{R}^k$ iff (t, X) is normal on $\mathbb{R}$ for all $t \in \mathbb{R}^k$.*

Proof. "$\Longrightarrow$". The c.f. of real-valued random variable (t, X) is

$$f(\lambda) = \mathbb{E}e^{i\lambda(t,X)} = \mathbb{E}e^{i(\lambda t,X)} = e^{-\frac{1}{2}(C\lambda t,\lambda t)} = e^{-\frac{1}{2}\lambda^2(Ct,t)}$$

which means that $(t, X) \sim \mathcal{N}(0, (Ct, t))$.

"$\Longleftarrow$". If (t, X) is normal then

$$\mathbb{E}e^{i(t,X)} = e^{-\frac{1}{2}(Ct,t)}$$

because the variance of (t, X) is (Ct, t).

□

Lemma 23 *Let $Z = (X, Y)$ where $X = (X_1, \ldots, X_i)$ and $Y = (Y_1, \ldots, Y_j)$ and suppose that Z is normal on $\mathbb{R}^{i+j}$. Then X and Y are independent iff $\mathrm{Cov}(X_m, Y_n) = 0$ for all m, n.*

Proof. One way is obvious. The other way around, suppose that

$$C = \mathrm{Cov}(Z) = \begin{bmatrix} D & 0 \\ 0 & F \end{bmatrix}.$$

Then the c.f. of Z is

$$\mathbb{E}e^{i(t,Z)} = e^{-\frac{1}{2}(Ct,t)} = e^{-\frac{1}{2}(Dt_1,t_1)-\frac{1}{2}(Ft_2,t_2)} = \mathbb{E}e^{i(t_1,X)}\mathbb{E}e^{i(t_2,Y)},$$

where $t = (t_1, t_2)$. By uniqueness, X and Y are independent.

□

Lemma 24 *(Continuous Mapping.) Suppose that $\mathbb{P}_n \to \mathbb{P}$ on X and $G : X \to Y$ is a continuous map. Then $\mathbb{P}_n \circ G^{-1} \to \mathbb{P} \circ G^{-1}$ on Y. In other words, if r.v. $Z_n \to Z$ weakly then $G(Z_n) \to G(Z)$ weakly.*

Proof. This is obvious, because for any $f \in C_b(Y)$, we have $f \circ G \in C_b(X)$ and therefore,

$$\mathbb{E}f(G(Z_n)) \to \mathbb{E}f(G(Z)).$$

□

Lemma 25 *If $\mathbb{P}_n \to \mathbb{P}$ on $\mathbb{R}^k$ and $\mathbb{Q}_n \to \mathbb{Q}$ on $\mathbb{R}^m$ then $\mathbb{P}_n \times \mathbb{Q}_n \to \mathbb{P} \times \mathbb{Q}$ on $\mathbb{R}^{k+m}$.*

Proof. By Fubini theorem, The c.f.

$$\int e^{i(t,x)} d\mathbb{P}_n \times \mathbb{Q}_n(x) = \int e^{i(t_1,x_1)} d\mathbb{P}_n \int e^{i(t_2,x_2)} d\mathbb{Q}_n \to \int e^{i(t_1,x_1)} d\mathbb{P} \int e^{i(t_2,x_2)} d\mathbb{Q} = \int e^{i(t,x)} d\mathbb{P} \times \mathbb{Q}.$$

By Lemma 19 it remains to show that $(\mathbb{P}_n \times \mathbb{Q}_n)$ is uniformly tight. By Theorem 21, since $\mathbb{P}_n \to \mathbb{P}$, $(\mathbb{P}_n)$ is uniformly tight. Therefore, there exists a compact K on $\mathbb{R}^k$ such that $\mathbb{P}_n(K) > 1 - \varepsilon$. Similarly, for some compact K' on $\mathbb{R}^m$, $\mathbb{Q}_n(K') > 1 - \varepsilon$. We have,

$$\mathbb{P}_n \times \mathbb{Q}_n(K \times K') > 1 - 2\varepsilon$$

and $K \times K'$ is a compact on $\mathbb{R}^{k+m}$.

□

Corollary 1 *If $\mathbb{P}_n \to \mathbb{P}$ and $\mathbb{Q}_n \to \mathbb{Q}$ both on $\mathbb{R}^k$ then $\mathbb{P}_n * \mathbb{Q}_n \to \mathbb{P} * \mathbb{Q}$.*

Proof. Since a function $G : \mathbb{R}^{k+k} \to \mathbb{R}^k$ given by $G(x, y) = x + y$ is continuous, by continuous mapping lemma,

$$\mathbb{P}_n * \mathbb{Q}_n = (\mathbb{P}_n \times \mathbb{Q}_n) \circ G^{-1} \to (\mathbb{P} \times \mathbb{Q}) \circ G^{-1} = \mathbb{P} * \mathbb{Q}.$$

□

Section 11

Lindeberg's CLT. Levy's Equivalence Theorem. Three Series Theorem.

Instead of considering i.i.d. sequences, for each $n \geq 1$ we will consider a vector $(X_1^n, \ldots, X_n^n)$ of independent r.v., not necessarily identically distributed. This setting is called *triangular arrays* because the entire vector may change with n.

Theorem 25 *Consider a vector $(X_i^n)_{1 \leq i \leq n}$ of independent r.v.s such that*

$$\mathbb{E}X_i^n = 0, \;\; \mathrm{Var}(S_n) = \sum_{i \leq n} \mathbb{E}(X_i^n)^2 = 1.$$

Suppose that the following Lindeberg's condition *is satisfied:*

$$\sum_{i=1}^{n} \mathbb{E}(X_i^n)^2 \mathrm{I}(|X_i^n| > \varepsilon) \to 0 \quad \text{as } n \to \infty \;\; \text{for all } \varepsilon > 0. \tag{11.0.1}$$

Then $\mathcal{L}\Big(\sum_{i \leq n} X_i^n\Big) \to \mathcal{N}(0,1)$.

Proof. First of all, $\Big\{\mathcal{L}\Big(\sum_{i \leq n} X_i^n\Big)\Big\}$ is uniformly tight, because by Chebyshev's inequality

$$\mathbb{P}\Big(\Big|\sum_{i \leq n} X_i^n\Big| > M\Big) \leq \frac{1}{M^2} \leq \varepsilon$$

for large enough M. It remains to show that the characteristic function of S_n coverges to $e^{-\frac{\lambda^2}{2}}$. For simplicity of notations let us omit the upper index n and write X_i instead of X_i^n. Since,

$$\mathbb{E}e^{i\lambda S_n} = \prod_{i \leq n} \mathbb{E}e^{i\lambda X_i}$$

it is enough to show that

$$\log \mathbb{E}e^{i\lambda S_n} = \sum \log\Big(1 + \big(\mathbb{E}e^{i\lambda X_i} - 1\big)\Big) \to -\frac{\lambda^2}{2}. \tag{11.0.2}$$

It is an easy exercise to prove, by induction on m, that for any $a \in \mathbb{R}$,

$$\Big|e^{ia} - \sum_{k \leq m} \frac{(ia)^k}{k!}\Big| \leq \frac{|a|^{m+1}}{(m+1)!}. \tag{11.0.3}$$

(Just integrate this inequality to make the induction step.) Using this for $m = 1$,

$$\left|\mathbb{E}e^{i\lambda X_i} - 1\right| = \left|\mathbb{E}e^{i\lambda X_i} - 1 - i\lambda\mathbb{E}X_i\right| \le \frac{\lambda^2}{2}\mathbb{E}X_i^2 \le \frac{\lambda^2}{2}\varepsilon^2 + \frac{\lambda^2}{2}\mathbb{E}X_i^2\mathrm{I}(|X_i| > \varepsilon) \le \lambda^2\varepsilon^2 \le \frac{1}{2} \tag{11.0.4}$$

for large n by (11.0.1) and for small enough ε. Using the expansion of $\log(1+z)$ it is easy to check that

$$\left|\log(1+z) - z\right| \le |z|^2 \quad \text{for} \quad |z| \le \frac{1}{2}$$

and, therefore, we can write

$$\begin{aligned}\sum_{i\le n}\left|\log\bigl(1 + \bigl(\mathbb{E}e^{i\lambda X_i} - 1\bigr)\bigr) - \bigl(\mathbb{E}e^{i\lambda X_i} - 1\bigr)\right| &\le \sum_{i\le n}\bigl|\mathbb{E}e^{i\lambda X_i} - 1\bigr|^2 \le \sum_{i\le n}\frac{\lambda^4}{4}\bigl(\mathbb{E}X_i^2\bigr)^2 \\ &\le \frac{\lambda^4}{4}\bigl(\max_{i\le n}\mathbb{E}X_i^2\bigr)\sum_{i\le n}\mathbb{E}X_i^2 = \frac{\lambda^4}{4}\max_{i\le n}\mathbb{E}X_i^2 \to 0\end{aligned}$$

because, as in (11.0.4),

$$\mathbb{E}X_i^2 \le \varepsilon^2 + \mathbb{E}X_i^2\mathrm{I}(|X_i| > \varepsilon) \to 0$$

for large n by (11.0.1) and for $\varepsilon \to 0$. Finally, to show (11.0.2) it remains to show that

$$\sum_{i\le n}\bigl(\mathbb{E}e^{i\lambda X_i} - 1\bigr) \to -\frac{\lambda^2}{2}.$$

Using (11.0.3) for $m = 1$, on the event $\bigl\{|X_i| > \varepsilon\bigr\}$,

$$\left|e^{i\lambda X_i} - 1 - i\lambda X_i\right|\mathrm{I}\bigl(|X_i| > \varepsilon\bigr) \le \frac{\lambda^2}{2}X_i^2\mathrm{I}\bigl((|X_i| > \varepsilon\bigr)$$

and, therefore,

$$\left|e^{i\lambda X_i} - 1 - i\lambda X_i + \frac{\lambda^2}{2}X_i^2\right|\mathrm{I}\bigl(|X_i| > \varepsilon\bigr) \le \lambda^2 X_i^2\mathrm{I}\bigl(|X_i| > \varepsilon\bigr).$$

Using (11.0.3) for $m = 2$, on the event $\bigl\{|X_i| \le \varepsilon\bigr\}$,

$$\left|e^{i\lambda X_i} - 1 - i\lambda X_i + \frac{\lambda^2}{2}X_i^2\right|\mathrm{I}\bigl(|X_i| \le \varepsilon\bigr) \le \frac{\lambda^3}{6}|X_i|^3\mathrm{I}\bigl(|X_i| \le \varepsilon\bigr) \le \frac{\lambda^3\varepsilon}{6}X_i^2.$$

Combining the last two equations and using that $\mathbb{E}X_i = 0$,

$$\left|\mathbb{E}e^{i\lambda X_i} - 1 + \frac{\lambda^2}{2}\mathbb{E}X_i^2\right| \le \lambda^2\mathbb{E}X_i^2\mathrm{I}\bigl(|X_i| > \varepsilon\bigr) + \frac{\lambda^3\varepsilon}{6}\mathbb{E}X_i^2.$$

Finally,

$$\begin{aligned}\left|\sum_{i\le n}\bigl(\mathbb{E}e^{i\lambda X_i} - 1\bigr) + \frac{\lambda^2}{2}\right| &= \left|\sum_{i\le n}\bigl(\mathbb{E}e^{i\lambda X_i} - 1\bigr) + \frac{\lambda^2}{2}\sum_{i\le n}\mathbb{E}X_i^2\right| \\ &\le \frac{\lambda^3\varepsilon}{6} + \lambda^2\sum\mathbb{E}X_i^2\mathrm{I}(|X_i| > \varepsilon) \to 0\end{aligned}$$

as $n \to \infty$ (using Lindeberg's condition) and $\varepsilon \to 0$.

□

Lemma 26 *If $\mathbb{P}, \mathbb{Q}$ are distributions on $\mathbb{R}$ such that $\mathbb{P} * \mathbb{Q} = \mathbb{P}$ then $\mathbb{Q}(\{0\}) = 1$.*

Proof. Let us define

$$f_P(t) = \int e^{itx} d\mathbb{P}(x), \quad f_Q(t) = \int e^{itx} d\mathbb{Q}(x).$$

The condition $\mathbb{P} * \mathbb{Q} = \mathbb{P}$ implies that $f_P(t) f_Q(t) = f_P(t)$. Since $f_P(0) = 1$ and $f_P(t)$ is continuous, for small enough $|t| \le \varepsilon$ we have $|f_P(t)| > 0$ and, as a result, $f_Q(t) = 1$. Since

$$f_Q(t) = \int \cos(tx) d\mathbb{Q}(x) + i \int \sin(tx) d\mathbb{Q}(x)$$

for $|t| \le \varepsilon$ this implies that $\int \cos(tx) d\mathbb{Q}(x) = 1$ and since $\cos(s) \le 1$ this can happen only if

$$\mathbb{Q}\Big(\{ x : xt = 0 \bmod 2\pi \} \Big) = 1 \text{ for all } |t| \le \varepsilon.$$

Take s, t such that $|s|, |t| \le \varepsilon$ and s/t is irrational. For x to be in the support of $\mathbb{Q}$ we must have $xs = 2\pi k$ and $xt = 2\pi m$ for some integer k, m. This can happen only if $x = 0$.

□

Theorem 26 *(Levy's equivalence) If (X_i) is a sequence of independent r.v. then $\sum_{i \ge 1} X_i$ converges a.s. iff in probability iff in law.*

Proof. We already proved (a Kolmogorov's theorem) that convergence in probability implies a.s. convergence. It remains to prove only that convergence in law implies convergence in probability.

Suppose that $\mathcal{L}(S_n) \to \mathbb{P}$. Convergence in law implies that $\{\mathcal{L}(S_n)\}$ is uniformly tight which easily implies that $\{\mathcal{L}(S_n - S_k)\}_{n,k \le 1}$ is uniformly tight. This will imply that for any $\varepsilon > 0$

$$\mathbb{P}(|S_n - S_k| > \varepsilon) < \varepsilon \tag{11.0.5}$$

for $n \ge k \ge N$ for large enough N. Suppose not. Then there exists $\varepsilon > 0$ and sequences $(n(l))$ and $(n'(l))$ such that $n(l) \le n'(l)$ and

$$\mathbb{P}(|S_{n'(l)} - S_{n(l)}| > \varepsilon) \ge \varepsilon.$$

Let us denote $Y_l = S_{n'(l)} - S_{n(l)}$. Since $\{\mathcal{L}(Y_l)\}$ is uniformly tight, by selection theorem, there exists a subsequence $(l(r))$ such that $\mathcal{L}(Y_{l(r)}) \to \mathbb{Q}$. Since

$$S_{n'(l(r))} = S_{n(l(r))} + Y_{l(r)} \quad \text{and} \quad \mathcal{L}(S_{n'(l(r))}) = \mathcal{L}(S_{n(l(r))}) * \mathcal{L}(Y_{l(r)})$$

letting $r \to \infty$ we get that $\mathbb{P} = \mathbb{P} * \mathbb{Q}$. By the above Lemma, $\mathbb{Q}(\{0\}) = 1$, which implies that $\mathbb{P}(|Y_{l(r)}| > \varepsilon) \le \varepsilon$ for large r - a contradiction. Once (11.0.5) is proved, by Borel-Cantelli lemma we can choose a.s. converging subsequence as in Kolmogorov's theorem, and then by (11.0.5) S_n converges in probability to the same limit.

□

Theorem 27 *(Three series theorem) Let $(X_i)_{i \ge 1}$ be a sequence of independent r.v. and let*

$$Z_i = X_i \mathrm{I}(|X_i| \le 1).$$

Then $\sum_{i \ge 1} X_i$ converges iff

1. $\sum_{i \ge 1} \mathbb{P}(|X_i| > 1) < \infty$,
2. $\sum_{i \ge 1} \mathbb{E} Z_i$ *converges,*
3. $\sum_{i \ge 1} \mathrm{Var}(Z_i) < \infty$.

Proof. "$\Longleftarrow$". Suppose 1 - 3 hold. Since

$$\sum_{i \ge 1} \mathbb{P}(|X_i| > 1) = \sum_{i \ge 1} \mathbb{P}(X_i \ne Z_i) < \infty$$

by Borel-Cantelli lemma $\mathbb{P}(\{X_i \neq Z_i\} \text{ i.o.}) = 0$ which means that $\sum_{i\geq 1} X_i$ converges iff $\sum_{i\geq 1} Z_i$ converges. By 2, it is enough to show that $\sum_{i\geq 1}(Z_i - \mathbb{E}Z_i)$ converges, but this follows from Theorem 13 by 3.

"$\Longrightarrow$". If $\sum_{i\geq 1} X_i$ converges a.s.,

$$\mathbb{P}(\{|X_i| > 1\} \text{ i.o}) = 0$$

and since (X_i) are independent, by Borel-Cantelli,

$$\sum_{i\geq 1} \mathbb{P}(|X_i| > 1) < \infty.$$

If $\sum_{i\geq 1} X_i$ converges then, obviously, $\sum_{i\geq 1} Z_i$ converges. Let

$$S_{mn} = \sum_{m\leq k\leq n} Z_k.$$

Since $S_{mn} \to 0$ as $m, n \to \infty$, $\mathbb{P}(|S_{mn}| > \delta) \leq \delta$ for any $\delta > 0$ for m, n large enough. Suppose that $\sum_{i\geq 1} \mathrm{Var}(Z_i) = \infty$. Then

$$\sigma_{mn}^2 = \mathrm{Var}(S_{mn}) = \sum_{m\leq k\leq n} \mathrm{Var}(Z_k) \to \infty$$

as $n \to \infty$ for any fixed m. Intuitively, this should not happen: $S_{mn} \to 0$ in probability but their variance goes to infinity. In principle, one can construct such sequence of random variables but in our case it will be ruled out by Lindeberg's CLT. Because $\sigma_{mn} \to \infty$, Lindeberg's theorem will imply that,

$$T_{mn} = \frac{S_{mn} - \mathbb{E}S_{mn}}{\sigma_{mn}} = \sum_{m\leq k\leq n} \frac{Z_k - \mathbb{E}Z_k}{\sigma_{mn}} \to \mathcal{N}(0,1),$$

if $m, n \to \infty$ and $\sigma_{mn}^2 \to \infty$. We only need to check that

$$\sum_{m\leq k\leq n} \mathbb{E}\Big(\frac{Z_k - \mathbb{E}Z_k}{\sigma_{mn}}\Big)^2 \mathrm{I}\Big(\Big|\frac{Z_k - \mathbb{E}Z_k}{\sigma_{mn}}\Big| > \varepsilon\Big) \to 0$$

as $m, n, n-m \to \infty$. Since $|Z_k - \mathbb{E}Z_k| < 2$ and $\sigma_{mn} \to \infty$, the event in the indicator does not occur for large m, n and, therefore, the sum is equal to 0. Next, since $\mathbb{P}(|S_{mn}| \leq \delta) \geq 1 - \delta$ we get

$$P\Big(\Big|T_{mn} + \frac{\mathbb{E}S_{mn}}{\sigma_{mn}}\Big| \leq \frac{\delta}{\sigma_{mn}}\Big) \geq 1 - \delta.$$

But this is impossible, since T_{mn} is approximately standard normal and standard normal distribution does not concentrate near any constant. We proved that $\sum_{i\geq 1} \mathrm{Var}(Z_i) < \infty$. By Kolmogorov's SLLN this implies that $\sum_{i\geq 1}(Z_i - \mathbb{E}Z_i)$ converges and, therefore, $\sum_{i\geq 1} \mathbb{E}Z_i$ converges.

□

Section 12

Levy's Continuity Theorem. Poisson Approximation. Conditional Expectation.

Let us start with the following bound.

Lemma 27 *Let X be a real-valued r.v. with distribution $\mathbb{P}$ and let*

$$f(t) = \mathbb{E}e^{itX} = \int e^{itx} d\mathbb{P}(x).$$

Then,

$$\mathbb{P}\Big(|X| > \frac{1}{u}\Big) \le \frac{7}{u}\int_0^u (1 - \mathbf{Re} f(t))dt.$$

Proof. Since

$$\mathbf{Re} f(t) = \int \cos tx d\mathbb{P}(x)$$

we have

$$\begin{aligned}
\frac{1}{u}\int_0^u \int_{\mathbb{R}} (1-\cos tx) d\mathbb{P}(x)dt &= \frac{1}{u}\int_{\mathbb{R}} \int_0^u (1-\cos tx)dt d\mathbb{P}(x) \\
&= \int_{\mathbb{R}} \Big(1 - \frac{\sin xu}{xu}\Big) d\mathbb{P}(x) \\
&\ge \int_{|xu|\ge 1} \Big(1 - \frac{\sin xu}{xu}\Big) d\mathbb{P}(x) \\
\Big\{\text{since } \frac{\sin y}{y} < \frac{\sin 1}{1} \text{ if } y > 1\Big\} &\ge (1-\sin 1)\int_{|xu|\ge 1} 1 d\mathbb{P}(x) \ge \frac{1}{7}\mathbb{P}\Big(|X| \ge \frac{1}{u}\Big).
\end{aligned}$$

□

Theorem 28 *(Levy continuity) Let (X_n) be a sequence of r.v. on $\mathbb{R}^k$. Suppose that*

$$f_n(t) = \mathbb{E}e^{i(t,X_n)} \to f(t)$$

and $f(t)$ is continuous at 0 along each axis. Then there exists a probability distribution $\mathbb{P}$ such that

$$f(t) = \int e^{i(t,x)} d\mathbb{P}(x)$$

and $\mathcal{L}(X_n) \to \mathbb{P}$.

Proof. By Lemma 19 we only need to show that $\{\mathcal{L}(X_n)\}$ is uniformly tight. If we denote

$$X_n = (X_{n,1}, \ldots, X_{n,k})$$

then the c.f.s along the i^{th} coordinate:

$$f_n^i(t_i) := f_n(0, \ldots, t_i, 0, \ldots 0) = \mathbb{E} e^{it_i X_{n,i}} \to f(0, \ldots, t_i, \ldots 0) =: f^i(t_i).$$

Since $f_n(0) = 1$ and, therefore, $f(0) = 1$, for any $\varepsilon > 0$ we can find $\delta > 0$ such that for all $i \le k$

$$|f^i(t_i) - 1| \le \varepsilon \quad \text{if} \quad |t_i| \le \delta.$$

This implies that for large enough n

$$|f_n^i(t_i) - 1| \le 2\varepsilon \quad \text{if} \quad |t_i| \le \delta.$$

Using previous Lemma,

$$\mathbb{P}\Big(|X_{n,i}| > \frac{1}{\delta}\Big) \le \frac{7}{\delta} \int_0^\delta \Big(1 - \mathbf{Re} f_n^i(t_i)\Big) dt_i \le \frac{7}{\delta} \int_0^\delta \big|1 - f_n^i(t_i)\big| \, dt_i \le 7 \cdot 2\varepsilon.$$

The union bound implies that

$$\mathbb{P}\Big(|X_n| > \frac{\sqrt{k}}{\delta}\Big) \le 14 k \varepsilon$$

and $\{\mathcal{L}(X_n)\}_{n \ge 1}$ is uniformly tight.

□

CLT describes how sums of independent r.v.s are approximated by normal distribution. We will now give a simple example of a different approximation. Consider independent Bernoulli random variables $X_i^n \sim B(p_i^n)$ for $i \le n$, i.e. $\mathbb{P}(X_i^n = 1) = p_i^n$ and $\mathbb{P}(X_i^n = 0) = 1 - p_i^n$. If $p_i^n = p > 0$ then by CLT

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to \mathcal{N}(0, 1).$$

However, if $p = p_i^n \to 0$ fast enough then, for example, the Lindeberg conditions will be violated. It is well-known that if $p_i^n = p_n$ and $np_n \to \lambda$ then S_n has approximately Poisson distribution Π_λ with p.f.

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda} \text{ for } k = 0, 1, 2, \ldots$$

Here is a version of this result.

Theorem 29 *Consider independent $X_i \sim B(p_i)$ for $i \le n$ and let*

$$S_n = X_1 + \ldots + X_n \text{ and } \lambda = p_1 + \ldots + p_n.$$

Then for any subset of integers $B \subseteq \mathbb{Z}$,

$$|\mathbb{P}(S_n \in B) - \Pi_\lambda(B)| \le \sum_{i \le n} p_i^2.$$

Proof. The proof is based on the construction on "one probability space". Let us construct Bernoulli r.v. $X_i \sim B(p_i)$ and Poisson r.v. $X_i^* \sim \Pi_{p_i}$ on the same probability space as follows. Let us consider a probability space $([0,1], \mathcal{B}, \lambda)$ with Lebesque measure λ. Define

$$X_i = X_i(x) = \begin{cases} 0, & 0 \le x \le 1 - p_i, \\ 1, & 1 - p_i < x \le 1. \end{cases}$$

Clearly, $X_i \sim B(p_i)$. Let us construct X_i^* as follows. If for $k \ge 0$ we define

$$c_k = \sum_{0 \le l \le k} \frac{(p_i)^l}{l!} e^{-p_i}$$

then

$$X_i = X_i(x) = \begin{cases} 0, & 0 \le x \le c_0, \\ 1, & c_0 < x \le c_1, \\ 2, & c_1 < x \le c_2, \\ \dots \end{cases}$$

Clearly, $X_i^* \sim \Pi_{p_i}$. When $X_i \neq X_i^*$? Since $1 - p_j \le e^{-p_j} = c_0$, this can only happen for

$$1 - p_i < x \le c_0 \quad \text{and} \quad c_1 < x \le 1,$$

i.e.

$$\mathbb{P}(X_j \neq X_j^*) = e^{p_j} - (1 - p_j) + (1 - e^{-p_j} - p_j e^{-p_j}) = p_j(1 - e^{-p_j}) \le p_j^2$$

We construct pairs (X_i, X_i^*) on separate coordinates of a product space, thus, making them independent fo $i \le n$. It is well-known that $\sum_{i \le n} X_i^* \sim \Pi_\lambda$ and, finally, we get

$$\mathbb{P}(S_n \neq S_n^*) \le \sum_{j \le n} \mathbb{P}(X_j \neq X_j^*) \le \sum_{j \le n} p_j^2.$$

□

Conditional expectation. Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable such that $\mathbb{E}|X| < \infty$. Let $\mathcal{A}$ be a σ-subalgebra of $\mathcal{B}$, $\mathcal{A} \subseteq \mathcal{B}$.

Definition. $Y = \mathbb{E}(X|\mathcal{A})$ is called *conditional expectation* of X given $\mathcal{A}$ if

1. $Y : \Omega \to \mathbb{R}$ is measurable on $\mathcal{A}$, i.e. if B is a Borel set on $\mathbb{R}$ then $Y^{-1}(B) \in \mathcal{A}$.
2. For any set $A \in \mathcal{A}$ we have $\mathbb{E}X\mathrm{I}_A = \mathbb{E}Y\mathrm{I}_A$, where $I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$

Definition. If X, Z are random variables then conditional expectation of X given Z is defined by

$$Y = \mathbb{E}(X|Z) = \mathbb{E}(X|\sigma(Z)).$$

Since Y is measurable on $\sigma(Z)$, $Y = f(Z)$ for some measurable function f.

Properties of conditional expectation.

1. (Existence of conditional expectation.) Let us define

$$\mu(A) = \int_A X d\mathbb{P} \quad \text{for} \quad A \in \mathcal{A}.$$

$\mu(A)$ is a σ-additive signed measure on $\mathcal{A}$. Since X is integrable, if $\mathbb{P}(A) = 0$ then $\mu(A) = 0$ which means that μ is absolutely continuous w.r.t. $\mathbb{P}$. By Radon-Nikodym theorem, there exists $Y = \frac{d\mu}{d\mathbb{P}}$ measurable on $\mathcal{A}$ such that for $A \in \mathcal{A}$

$$\mu(A) = \int_A X d\mathbb{P} = \int_A Y d\mathbb{P}.$$

By definition $Y = \mathbb{E}(X|\mathcal{A})$.

2. (Uniqueness) Suppose there exists $Y' = \mathbb{E}(X|\mathcal{A})$ such that $\mathbb{P}(Y \neq Y') > 0$, i.e.

$$\mathbb{P}(Y > Y') > 0 \text{ or } \mathbb{P}(Y < Y') > 0.$$

Since both Y, Y' are measurable on $\mathcal{A}$ the set $A = \{Y > Y'\} \in \mathcal{A}$. One one hand, $\mathbb{E}(Y - Y')\mathrm{I}_A > 0$. On the other hand,

$$\mathbb{E}(Y - Y')\mathrm{I}_A = \mathbb{E}X\mathrm{I}_A - \mathbb{E}X\mathrm{I}_A = 0$$

- a contradiction.

3. $\mathbb{E}(cX + Y|\mathcal{A}) = c\mathbb{E}(X|\mathcal{A}) + \mathbb{E}(Y|\mathcal{A})$.

4. If σ-algebras $\mathcal{C} \subseteq \mathcal{A} \subseteq \mathcal{B}$ then

$$\mathbb{E}(\mathbb{E}(X|\mathcal{A})|\mathcal{C}) = \mathbb{E}(X|\mathcal{C}).$$

Consider a set $C \in \mathcal{C} \subseteq \mathcal{A}$. Then

$$\mathbb{E}\mathrm{I}_C(\mathbb{E}(\mathbb{E}(X|\mathcal{A})|\mathcal{C})) = \mathbb{E}\mathrm{I}_C\mathbb{E}(X|\mathcal{A}) = \mathbb{E}\mathrm{I}_C X \quad \text{and} \quad \mathbb{E}\mathrm{I}_C(\mathbb{E}(X|\mathcal{C})) = \mathbb{E}X\mathrm{I}_C.$$

We conclude by uniqueness.

5. $\mathbb{E}(X|\mathcal{B}) = X$, $\mathbb{E}(X|\{\emptyset, \Omega\}) = \mathbb{E}X$, $\mathbb{E}(X|\mathcal{A}) = \mathbb{E}X$ if X is independent of $\mathcal{A}$.

6. If $X \leq Z$ then $\mathbb{E}(X|\mathcal{A}) \leq \mathbb{E}(Z|\mathcal{A})$ a.s.; proof is similar to proof of uniqueness.

7. (Monotone convergence) If $\mathbb{E}|X_n| < \infty, \mathbb{E}|X| < \infty$ and $X_n \uparrow X$ then $\mathbb{E}(X_n|\mathcal{A}) \uparrow \mathbb{E}(X|\mathcal{A})$. Since

$$\mathbb{E}(X_n|\mathcal{A}) \leq \mathbb{E}(X_{n+1}|\mathcal{A}) \leq \mathbb{E}(X|\mathcal{A})$$

there exists a limit

$$g = \lim_{n\to\infty} \mathbb{E}(X_n|\mathcal{A}) \leq \mathbb{E}(X|\mathcal{A}).$$

Since $\mathbb{E}(X_n|\mathcal{A})$ are measurable on $\mathcal{A}$, so is $g = \lim \mathbb{E}(X_n|\mathcal{A})$. It remains to check that

$$\text{for any set } A \in \mathcal{A}, \ \ \mathbb{E}g\mathrm{I}_A = \mathbb{E}X\mathrm{I}_A.$$

Since $X_n\mathrm{I}_A \uparrow X\mathrm{I}_A$ and $\mathbb{E}(X_n|\mathcal{A})\mathrm{I}_A \uparrow g\mathrm{I}_A$, by monotone convergence theorem,

$$\mathbb{E}X_n\mathrm{I}_A \uparrow \mathbb{E}X\mathrm{I}_A \quad \text{and} \quad \mathbb{E}\mathrm{I}_A\mathbb{E}(X_n|\mathcal{A}) \uparrow \mathbb{E}g\mathrm{I}_A.$$

But since $\mathbb{E}\mathrm{I}_A\mathbb{E}(X_n|\mathcal{A}) = \mathbb{E}X_n\mathrm{I}_A$ this implies that $\mathbb{E}g\mathrm{I}_A = \mathbb{E}X\mathrm{I}_A$ and, therefore, $g = \mathbb{E}(X|\mathcal{A})$ a.s.

8. (Dominated convergence) If $|X_n| \leq Y, \mathbb{E}Y < \infty$, and $X_n \to X$ then

$$\lim \mathbb{E}(X_n|A) = \mathbb{E}(X|A).$$

We can write,

$$-Y \leq g_n = \inf_{m\geq n} X_m \leq X_n \leq h_n = \sup_{m\geq n} X_m \leq Y.$$

Since

$$g_n \uparrow X, \ \ h_n \downarrow X, \ \ |g_n| \leq Y, |h_n| \leq Y$$

by monotone convergence

$$\mathbb{E}(g_n|\mathcal{A}) \uparrow \mathbb{E}(X|\mathcal{A}), \ \ \mathbb{E}(h_n|\mathcal{A}) \downarrow \mathbb{E}(X|\mathcal{A}) \Longrightarrow \mathbb{E}X_n|A) \to \mathbb{E}(X|A).$$

9. If $\mathbb{E}|X| < \infty, \mathbb{E}|XY| < \infty$ and Y is measurable on $\mathcal{A}$ then

$$\mathbb{E}(XY|\mathcal{A}) = Y\mathbb{E}(X|\mathcal{A}).$$

We can assume that $X, Y \geq 0$ by decomposing $X = X^+ - X^-, Y = Y^+ - Y^-$. Consider a sequence of simple functions

$$Y_n = \sum w_k \mathrm{I}_{C_k}, \;\; C_k \in \mathcal{A}$$

measurable on $\mathcal{A}$ such that $0 \leq Y_n \uparrow Y$. By monotone convergence theorem, it is enough to prove that

$$\mathbb{E}(X\mathrm{I}_{C_k}|\mathcal{A}) = \mathrm{I}_{C_k}\mathbb{E}(X|\mathcal{A}).$$

Take $B \in \mathcal{A}$. Since $BC_k \in \mathcal{A}$,

$$\mathbb{E}\mathrm{I}_B\mathrm{I}_{C_k}\mathbb{E}(X|\mathcal{A}) = \mathbb{E}\mathrm{I}_{BC_k}\mathbb{E}(X|\mathcal{A}) = \mathbb{E}X\mathrm{I}_{BC_k} = \mathbb{E}(X\mathrm{I}_{C_k})\mathrm{I}_B.$$

10. (Jensen's inequality) If $f : \mathbb{R} \to \mathbb{R}$ is convex then

$$f(\mathbb{E}(X|\mathcal{A})) \leq \mathbb{E}(f(X|\mathcal{A})).$$

By convexity,

$$f(X) - f(\mathbb{E}(X|\mathcal{A})) \geq \partial f(\mathbb{E}(X|\mathcal{A}))(X - \mathbb{E}(X|\mathcal{A})).$$

Taking condition expectation of both sides,

$$\mathbb{E}(f(X)|\mathcal{A}) - f(\mathbb{E}(X|\mathcal{A})) \geq \partial f(\mathbb{E}(X|\mathcal{A}))(\mathbb{E}(X|\mathcal{A}) - \mathbb{E}(X|\mathcal{A})) = 0.$$

□

Section 13

Martingales. Doob's Decomposition. Uniform Integrability.

Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space and let $(T, \leq)$ be a linearly ordered set. Consider a family of σ-algebras $\mathcal{B}_t, t \in T$ such that for $t \leq u$, $\mathcal{B}_t \subseteq \mathcal{B}_u \subseteq \mathcal{B}$.

Definition. A family $(X_t, \mathcal{B}_t)_{t\in T}$ is called a *martingale* if

1. $X_t : \Omega \to \mathbb{R}$ is measurable w.r.t. $\mathcal{B}_t$; in other words, X_t is *adapted* to $\mathcal{B}_t$.
2. $\mathbb{E}|X_t| < \infty$.
3. $\mathbb{E}(X_u|\mathcal{B}_t) = X_t$ for $t \leq u$.

If the last equality is replaced by $\mathbb{E}(X_u|\mathcal{B}_t) \leq X_t$ then the process is called a *supermartingale* and if $\mathbb{E}(X_u|\mathcal{B}_t) \geq X_t$ then it is called a *submartingale.*

Examples.

1. Consider a sequence $(X_n)_{n\geq 1}$ of independent random variables such that $\mathbb{E}X_i = 0$ and let $S_n = \sum_{i\leq n} X_i$. If $\mathcal{B}_n = \sigma(X_1, \ldots, X_n)$ is a σ-algebra generated by the first n r.v.s then $(S_n, \mathcal{B}_n)_{n\geq 1}$ is a martingale since

$$\mathbb{E}(S_{n+1}|\mathcal{B}_n) = \mathbb{E}(X_{n+1} + S_n|\mathcal{B}_n) = 0 + S_n = S_n.$$

2. Consider a sequence of σ-algebras

$$\ldots \subseteq \mathcal{B}_m \subseteq \mathcal{B}_n \subseteq \ldots \subseteq \mathcal{B}$$

and a r.v. X on $\mathcal{B}$ and let $X_n = \mathbb{E}(X|\mathcal{B}_n)$. Then $(X_n, \mathcal{B}_n)$ is a martingale since for $m < n$

$$\mathbb{E}(X_n|\mathcal{B}_m) = \mathbb{E}(\mathbb{E}(X|\mathcal{B}_n)|\mathcal{B}_m) = \mathbb{E}(X|\mathcal{B}_m) = X_m.$$

Definition. If $(X_n, \mathcal{B}_n)$ is a martingale and for some r.v. X, $X_n = \mathbb{E}(X|\mathcal{B}_n)$, then the martingale is called *right-closable.* If $X_\infty = X, \mathcal{B}_\infty = \mathcal{B}$ then $(X_n, \mathcal{B}_n)_{n\leq\infty}$ is called *right-closed.*

3. Let $(X_i)_{i\geq 1}$ be i.i.d. and let $S_n = \sum_{i\leq n} X_i$. Let us take $T = \{\ldots, -2, -1\}$ and for $n \geq 1$ define

$$\mathcal{B}_{-n} = \sigma(S_n, S_{n+1}, \ldots) = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots).$$

Clearly, $\mathcal{B}_{-(n+1)} \subseteq \mathcal{B}_{-n}$. For $1 \leq k \leq n$, by symmetry,

$$\mathbb{E}(X_1|\mathcal{B}_{-n}) = \mathbb{E}(X_k|\mathcal{B}_{-n}).$$

Therefore,

$$S_n = \mathbb{E}(S_n|\mathcal{B}_{-n}) = \sum_{1\leq k\leq n} \mathbb{E}(X_k|\mathcal{B}_{-n}) = n\mathbb{E}(X_1|\mathcal{B}_{-n}) \Longrightarrow Z_{-n} := \frac{S_n}{n} = \mathbb{E}(X_1|\mathcal{B}_{-n}).$$

Thus, $(Z_{-n}, \mathcal{B}_{-n})_{-n\leq -1}$ is a right-closed martingale.

□

Lemma 28 *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Suppose that either one of two conditions holds:*

1. *$(X_t, \mathcal{B}_t)$ is a martingale,*
2. *$(X_t, \mathcal{B}_t)$ is a submartingale and f is increasing.*

Then $(f(X_t), \mathcal{B}_t)$ is a submartingale.

Proof. 1. For $t \leq u$, by Jensen's inequality,

$$f(X_t) = f(\mathbb{E}(X_u|\mathcal{B}_t)) \leq \mathbb{E}(f(X_u)|\mathcal{B}_t).$$

2. For $t \leq u$, since $X_t \leq \mathbb{E}(X_u|\mathcal{B}_t)$ and f is increasing,

$$f(X_t) \leq f(\mathbb{E}(X_u|\mathcal{B}_t)) \leq \mathbb{E}(f(X_u)|\mathcal{B}_t),$$

where the last step is again Jensen's inequality.

□

Theorem 30 *(Doob's decomposition) If $(X_n, \mathcal{B}_n)_{n\geq 0}$ is a submartingale then it can be uniquely decomposed*

$$X_n = Z_n + Y_n,$$

where $(Y_n, \mathcal{B}_n)$ is a martingale, $Z_0 = 0, Z_n \leq Z_{n+1}$ almost surely and Z_n is $\mathcal{B}_{n-1}$-measurable.

Proof. Let $D_n = X_n - X_{n-1}$ and

$$G_n = \mathbb{E}(D_n|\mathcal{B}_{n-1}) = \mathbb{E}(X_n|\mathcal{B}_{n-1}) - X_{n-1} \geq 0$$

by the definition of submartingale. Let,

$$H_n = D_n - G_n, \;\; Y_n = H_1 + \ldots + H_n, \;\; Z_n = G_1 + \cdots + G_n.$$

Since $G_n \geq 0$ a.s., $Z_n \leq Z_{n+1}$ and, by construction, Z_n is $\mathcal{B}_{n-1}$-measurable. We have,

$$\mathbb{E}(H_n|\mathcal{B}_{n-1}) = \mathbb{E}(D_n|\mathcal{B}_{n-1}) - G_n = 0$$

and, therefore, $\mathbb{E}(Y_n|\mathcal{B}_{n-1}) = Y_{n-1}$. Uniqueness follows by construction. Suppose that $X_n = Z_n + Y_n$ with all stated properties. First, since $Z_0 = 0$, $Y_0 = X_0$. By induction, given a unique decomposition up to $n-1$, we can write

$$Z_n = \mathbb{E}(Z_n|\mathcal{B}_{n-1}) = \mathbb{E}(X_n - Y_n|\mathcal{B}_{n-1}) = \mathbb{E}(X_n|\mathcal{B}_{n-1}) - Y_{n-1}$$

and $Y_n = X_n - Z_n$.

□

Definition. We say that $(X_n)_{n\geq 1}$ is uniformly integrable if

$$\sup_n \mathbb{E}|X_n| < \infty \quad \text{and} \quad \sup_n \mathbb{E}|X_n|\mathrm{I}(|X_n| > M) \to 0 \quad \text{as} \quad M \to \infty.$$

Lemma 29 *The following holds.*

1. *If $(X_n, \mathcal{B}_n)$ is a right-closable martingale then (X_n) is uniformly integrable.*
2. *If $(X_n, \mathcal{B}_n)_{n\leq\infty}$ is a submartingale then for any $a \in \mathbb{R}$, $(\max(X_n, a))$ is uniformly integrable.*

Proof. 1. If $X_n = \mathbb{E}(Y|\mathcal{B}_n)$ then

$$|X_n| = |\mathbb{E}(Y|\mathcal{B}_n)| \le \mathbb{E}(|Y||\mathcal{B}_n) \quad \text{and} \quad \mathbb{E}|X_n| \le \mathbb{E}|Y| < \infty.$$

Since $\{|X_n| > M\} \in \mathcal{B}_n$,

$$X_n \mathrm{I}(|X_n| > M) = \mathrm{I}(|X_n| > M)\mathbb{E}(Y|\mathcal{B}_n) = \mathbb{E}(Y\mathrm{I}(|X_n| > M)|\mathcal{B}_n)$$

and, therefore,

$$\begin{aligned} \mathbb{E}|X_n|\mathrm{I}(|X_n| > M) \le \mathbb{E}|Y|\mathrm{I}(|X_n| > M) &\le K\mathbb{P}(|X_n| > M) + \mathbb{E}|Y|\mathrm{I}(|Y| > K) \\ &\le K\frac{\mathbb{E}|X_n|}{M} + \mathbb{E}|Y|\mathrm{I}(|Y| > K) \le K\frac{\mathbb{E}|Y|}{M} + \mathbb{E}|Y|\mathrm{I}(|Y| > K). \end{aligned}$$

Letting $M \to \infty, K \to \infty$ proves that $\sup_n \mathbb{E}|X_n|\mathrm{I}(|X_n| > M) \to 0$ as $M \to \infty$.

2. Since $(X_n, \mathcal{B}_n)_{n \le \infty}$ is a submartingale, for $Y = X_\infty$ we have $X_n \le \mathbb{E}(Y|\mathcal{B}_n)$. Below we will use the following observation. Since a function $\max(a, x)$ is convex and increasing in x, by Jensen's inequality

$$\max(a, X_n) \le \mathbb{E}(\max(a, Y)|\mathcal{B}_n). \tag{13.0.1}$$

Since,

$$\big|\max(X_n, a)\big| \le |a| + X_n \mathrm{I}\big(X_n > |a|\big)$$

and $\{|X_n| > |a|\} \in \mathcal{B}_n$ we can write

$$\mathbb{E}\big|\max(X_n, a)\big| \le |a| + \mathbb{E}X_n\mathrm{I}\big(X_n > |a|\big) \le |a| + \mathbb{E}Y\mathrm{I}\big(X_n > |a|\big) \le |a| + \mathbb{E}|Y| < \infty.$$

If we take $M > |a|$ then

$$\begin{aligned} \mathbb{E}|\max(X_n, a)|\mathrm{I}(|\max(X_n, a)| > M) &= \mathbb{E}X_n\mathrm{I}(X_n > M) \le \mathbb{E}Y\mathrm{I}(X_n > M) \\ &\le K\mathbb{P}(X_n > M) + \mathbb{E}|Y|\mathrm{I}(|Y| > K) \\ &\le K\frac{\mathbb{E}\max(X_n, 0)}{M} + \mathbb{E}|Y|\mathrm{I}(|Y| > K) \\ \text{by (13.0.1)} \quad &\le K\frac{\mathbb{E}\max(Y, 0)}{M} + \mathbb{E}|Y|\mathrm{I}(|Y| > K). \end{aligned}$$

Letting $M \to \infty$ and $K \to \infty$ finishes the proof.

□

Uniform integrability plays an important role when studying the convergence of martingales. The following strengthening of the dominated convergence theorem will be useful.

Lemma 30 *Consider r.v.s (X_n) and X such that $\mathbb{E}|X_n| < \infty, \mathbb{E}|X| < \infty$. Then the following are equivalent:*

1. $\mathbb{E}|X_n - X| \to 0$,
2. *(X_n) is uniformly integrable and $X_n \to X$ in probability.*

Proof. 2$\Longrightarrow$1. We can write,

$$\begin{aligned} \mathbb{E}|X_n - X| &\le \varepsilon + \mathbb{E}|X_n - X|\mathrm{I}(|X_n - X| > \varepsilon) \\ &\le \varepsilon + 2K\mathbb{P}(|X_n - X| > \varepsilon) + 2\mathbb{E}|X_n|\mathrm{I}(|X_n| > K) + 2\mathbb{E}|X|\mathrm{I}(|X| > K) \\ &\le \varepsilon + 2K\mathbb{P}(|X_n - X| > \varepsilon) + 2\sup_n \mathbb{E}|X_n|\mathrm{I}(|X_n| > K) + 2\mathbb{E}|X|\mathrm{I}(|X| > K). \end{aligned}$$

Letting $n \to \infty$ and then $\varepsilon \to 0, K \to \infty$ proves the result.

1$\Longrightarrow$2. By Chebyshev's inequality,

$$\mathbb{P}(|X_n - X| > \varepsilon) \le \frac{1}{\varepsilon}\mathbb{E}|X_n - X| \to 0$$

as $n \to \infty$ so $X_n \to X$ in probability. To prove uniform integrability let us first show that for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathbb{P}(A) < \delta \Longrightarrow \mathbb{E}|X|\mathrm{I}_A < \varepsilon.$$

Suppose not. Then, for some $\varepsilon > 0$ one can find a sequence of events $A(n)$ such that

$$\mathbb{P}(A(n)) \le \frac{1}{2^n} \quad \text{and} \quad \mathbb{E}|X|\mathrm{I}_{A(n)} > \varepsilon.$$

Since $\sum_{n\ge 1} \mathbb{P}(A(n)) < \infty$, by Borel-Cantelli lemma, $\mathbb{P}(A(n) \text{ i.o.}) = 0$. This means that $|X|\mathrm{I}_{A(n)} \to 0$ almost surely and by the dominated convergence theorem $\mathbb{E}|X|\mathrm{I}_{A(n)} \to 0$ - a contradiction.

Given $\varepsilon > 0$, take δ as above and take $M > 0$ large enough so that for all $n \ge 1$

$$\mathbb{P}(|X_n| > M) \le \frac{\mathbb{E}|X_n|}{M} < \delta.$$

Then,

$$\mathbb{E}|X_n|\mathrm{I}(|X_n| > M) \le \mathbb{E}|X_n - X| + \mathbb{E}|X|\mathrm{I}(|X_n| > M) \le \mathbb{E}|X_n - X| + \varepsilon.$$

For large enough $n \ge n_0$, $\mathbb{E}|X_n - X| \le \varepsilon$ and, therefore,

$$\mathbb{E}|X_n|\mathrm{I}(|X_n| > M) \le 2\varepsilon.$$

We can also choose M large enough so that $\mathbb{E}|X_n|\mathrm{I}(|X_n| > M) \le 2\varepsilon$ for $n \le n_0$ and this finishes the proof. □

Section 14

Optional stopping. Inequalities for martingales.

Consider a sequence of σ-algebras $(\mathcal{B}_n)_{n\geq 0}$ such that $\mathcal{B}_n \subseteq \mathcal{B}_{n+1}$. Integer valued r.v. $\tau \in \{1, 2, \ldots\}$ is called a *stopping time* if $\{\tau \leq n\} \in \mathcal{B}_n$. Let us denote by $\mathcal{B}_\tau$ a σ-algebra of the events B such that

$$\{\tau \leq n\} \cap B \in \mathcal{B}_n, \ \forall n \geq 1.$$

If (X_n) is adapted to $(\mathcal{B}_n)$ then random variables such as X_τ or $\sum_{k=1}^{\tau} X_k$ are measurable on $\mathcal{B}_\tau$. For example,

$$\{X_\tau \in A\} = \bigcup_{n\geq 1} \{\tau = n\} \cap \{X_n \in A\} = \bigcup_{n\geq 1} \Big(\{\tau \leq n\} \setminus \{\tau \leq n-1\} \bigcap \{X_n \in A\} \Big) \in \mathcal{B}_\tau.$$

Theorem 31 *(Optional stopping) Let $(X_n, \mathcal{B}_n)$ be a martingale and $\tau_1, \tau_2 < \infty$ be stopping times such that*

$$\mathbb{E}|X_{\tau_2}| < \infty, \ \lim_{n\to\infty} \mathbb{E}|X_n| \mathrm{I}(n \leq \tau_2) = 0. \tag{14.0.1}$$

Then on the event $\{\tau_1 \leq \tau_2\}$

$$\mathbb{E}(X_{\tau_2} | \mathcal{B}_{\tau_1}) = X_{\tau_1}.$$

More precisely, for any set $A \in \mathcal{B}_{\tau_1}$,

$$\mathbb{E} X_{\tau_2} \mathrm{I}_A \mathrm{I}(\tau_1 \leq \tau_2) = \mathbb{E} X_{\tau_1} \mathrm{I}_A \mathrm{I}(\tau_1 \leq \tau_2).$$

If $(X_n, \mathcal{B}_n)$ is a submartingale then equality is replaced by $\geq$.

Remark. If stopping times τ_1, τ_2 are bounded then (14.0.1) is satisfied. As the next example shows, without some control of the stopping times the statement is not true.

Example. Consider an i.i.d. sequence (X_n) such that

$$\mathbb{P}(X_n = \pm 2^n) = \frac{1}{2}.$$

If $\mathcal{B}_n = \sigma(X_1, \ldots, X_n)$ then $(S_n, \mathcal{B}_n)$ is a martingale. Let $\tau_1 = 1$ and $\tau_2 = \min\{k \geq 1, S_k > 0\}$. Clearly, $S_{\tau_2} = 2$ because if $\tau_2 = k$ then

$$S_{\tau_2} = S_k = -2 - 2^2 - \ldots - 2^{k-1} + 2^k = 2.$$

However,

$$2 = \mathbb{E}(S_{\tau_2} | \mathcal{B}_1) \neq S_{\tau_1} = X_1.$$

The second condition in (14.0.1) is violated since $\mathbb{P}(\tau_2 = n) = 2^{-n}$ and

$$\mathbb{E}|S_n|I(n \le \tau_2) = 2\mathbb{P}(\tau_2 = n) + (2^{n+1} - 2)\mathbb{P}(n+1 \le \tau_2) = 2 \not\to 0.$$

Proof of Theorem 31. Consider a set $A \in \mathcal{B}_{\tau_1}$. We have,

$$\begin{aligned}\mathbb{E}X_{\tau_2}\mathrm{I}_A\mathrm{I}(\tau_1 \le \tau_2) &= \sum_{n\ge 1}\mathbb{E}X_{\tau_2}\mathrm{I}\big(A \cap \{\tau_1 = n\}\big)\mathrm{I}(n \le \tau_2) \\ &\overset{(*)}{=} \sum_{n\ge 1}\mathbb{E}X_n\mathrm{I}\big(A \cap \{\tau_1 = n\}\big)\mathrm{I}(n \le \tau_2) = \mathbb{E}X_{\tau_1}\mathrm{I}_A\mathrm{I}(\tau_1 \le \tau_2).\end{aligned}$$

To prove (*) it is enough to prove that for $A_n = A \cap \{\tau_1 = n\} \in \mathcal{B}_n$,

$$\mathbb{E}X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2) = \mathbb{E}X_n\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2). \tag{14.0.2}$$

We can write

$$\begin{aligned}\mathbb{E}X_n\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2) &= \mathbb{E}X_n\mathrm{I}_{A_n}\mathrm{I}(\tau_2 = n) + \mathbb{E}X_n\mathrm{I}_{A_n}\mathrm{I}(n+1 \le \tau_2) \\ &= \mathbb{E}X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(\tau_2 = n) + \mathbb{E}X_n\mathrm{I}_{A_n}\mathrm{I}(n+1 \le \tau_2) \\ &\Big\{ \text{ since } \{n+1 \le \tau_2\} = \{\tau_2 \le n\}^c \in \mathcal{B}_n, \text{ by martingale property} \Big\} \\ &= \mathbb{E}X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(\tau_2 = n) + \mathbb{E}X_{n+1}\mathrm{I}_{A_n}\mathrm{I}(n+1 \le \tau_2) \\ &\{ \text{ by induction } \} = \sum_{n\le k<m}\mathbb{E}X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(\tau_2 = k) + \mathbb{E}X_m\mathrm{I}_{A_n}\mathrm{I}(m \le \tau_2) \\ &= \mathbb{E}X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2 < m) + \mathbb{E}X_m\mathrm{I}_{A_n}\mathrm{I}(m \le \tau_2).\end{aligned}$$

By (14.0.1), the last term

$$\big|\mathbb{E}X_m\mathrm{I}_{A_n}\mathrm{I}(m \le \tau_2)\big| \le \mathbb{E}|X_m|\mathrm{I}(m \le \tau_2) \to 0 \text{ as } m \to \infty.$$

Since

$$X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2 \le m) \to X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2) \text{ as } m \to \infty$$

and $\mathbb{E}|X_{\tau_2}| < \infty$, by dominated convergence theorem,

$$\mathbb{E}X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2 < m) \to \mathbb{E}X_{\tau_2}\mathrm{I}_{A_n}\mathrm{I}(n \le \tau_2).$$

This proves (14.0.2).

□

Theorem 32 *(Doob's inequality) If $(X_n, \mathcal{B}_n)$ is a submartingale then for $Y_n = \max_{1\le k\le n} X_k$ and $M > 0$*

$$\mathbb{P}\Big(Y_n \ge M\Big) \le \frac{1}{M}\mathbb{E}X_n I(Y_n \ge M) \le \frac{1}{M}\mathbb{E}X_n^+. \tag{14.0.3}$$

Proof. Define a stopping time

$$\tau_1 = \begin{cases} \min\{k : X_k \ge M, k \le n\} & \text{if such } k \text{ exists,} \\ n & \text{otherwise.}\end{cases}$$

Let $\tau_2 = n$ so that $\tau_1 \le \tau_2$. By Theorem 31,

$$\mathbb{E}(X_n|\mathcal{B}_{\tau_1}) = \mathbb{E}(X_{\tau_2}|\mathcal{B}_{\tau_1}) \ge X_{\tau_1}.$$

Let us apply this to the set $A = \{Y_n = \max_{1\le k\le n} X_n \ge M\}$ which belongs to $\mathcal{B}_{\tau_1}$ because

$$A \cap \{\tau_1 \le k\} = \Big\{\max_{1\le i\le k} X_i \ge M\Big\} \in \mathcal{B}_k.$$

On the event A, $X_{\tau_1} \geq M$ and, therefore,

$$\mathbb{E}X_n \mathrm{I}_A = \mathbb{E}X_{\tau_2}\mathrm{I}_A \geq \mathbb{E}X_{\tau_1}\mathrm{I}_A \geq M\mathbb{E}\mathrm{I}_A = M\mathbb{P}(A).$$

On the other hand, $\mathbb{E}X_n\mathrm{I}_A \leq \mathbb{E}X_n^+$ and this finishes the proof.

□

As a corollary we obtain the *second Kolmogorov's inequality*. If (X_i) are independent and $\mathbb{E}X_i = 0$ then $S_n = \sum_{1\leq i\leq n} X_i$ is a martingale and S_n^2 is a submartingale. Therefore,

$$\mathbb{P}\Big(\max_{1\leq k\leq n}|S_k| \geq M\Big) = \mathbb{P}\Big(\max_{1\leq k\leq n} S_k^2 \geq M^2\Big) \leq \frac{1}{M^2}\mathbb{E}S_n^2 = \frac{1}{M^2}\sum_{1\leq k\leq n}\mathrm{Var}(X_k).$$

Exercises.
1. Show that for any random variable Y, $\mathbb{E}|Y^p| = \int_0^\infty pt^{p-1}\mathbb{P}(|Y| \geq t)dt$.
2. Let X, Y be two non-negative random variables such that for every $t > 0$, $\mathbb{P}(Y \geq t) \leq t^{-1}\int XI(Y \geq t)d\mathbb{P}$. For any $p > 1$, $\|f\|_p = (\int |f|^p d\mathbb{P})^{1/p}$ and $1/p + 1/q = 1$, show that $\|Y\|_p \leq q\|X\|_p$.
3. Given a non-negative submartingale $(X_n, \mathcal{B}_n)$, let $X_n^* := \max_{j\leq n} X_j$ and $X^* := \max_{j\geq 1} X_j$. Prove that for any $p > 1$ and $1/p + 1/q = 1$, $\|X^*\|_p \leq q \sup_n \|X_n\|_p$. *Hint:* use exercise 2 and Doob's maximal inequality.

□

Doob's upcrossing inequality. Let $(X_n, \mathcal{B}_n)_{n\geq 1}$ be a submartingale. Given two real numbers $a < b$ we will define a sequence of stopping times (τ_n) when X_n is crossing a downward and b upward as in figure 14.1. Namely, we define

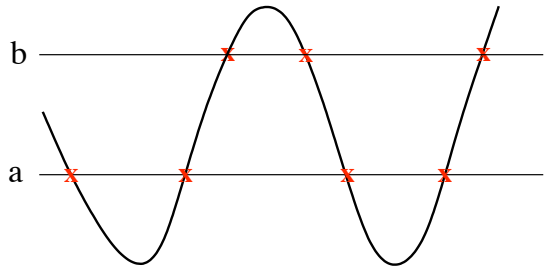


Figure 14.1: Stopping times of level crossings.

$$\tau_1 = \min\{n \geq 1, X_n \leq a\}, \;\; \tau_2 = \min\{n > \tau_2 : X_n \geq b\}$$

and, by induction, for $k \geq 2$

$$\tau_{2k-1} = \min\{n > \tau_{2k-2}, X_n \leq a\}, \;\; \tau_{2k} = \min\{n > 2k-1, X_n \geq b\}.$$

Define

$$\nu(a, b, n) = \max\{k : \tau_{2k} \leq n\}$$

- the number of upward crossings of $[a, b]$ before time n.

Theorem 33 *(Doob's upcrossing inequality) We have,*

$$\mathbb{E}\nu(a, b, n) \leq \frac{\mathbb{E}(X_n - a)^+}{b - a}. \tag{14.0.4}$$

Proof. Since $x \to (x - a)^+$ is increasing convex function, $Z_n = (X_n - a)^+$ is also a submartingale. Clearly,

$$\mu_X(a, b, n) = \nu_Z(0, b - a, n)$$

which means that it is enough to prove (14.0.4) for nonnegative submartingales. From now on we can assume that $0 \leq X_n$ and we would like to show that

$$\mathbb{E}\nu(0, b, n) \leq \frac{\mathbb{E}X_n}{b}.$$

Let us define a sequence of r.v.s

$$\eta_j = \begin{cases} 1, & \tau_{2k-1} < j \leq \tau_{2k} \text{ for some } k \\ 0, & \text{otherwise,} \end{cases}$$

i.e. η_j is the indicator of the event that at time j the process is crossing $[0, b]$ upward. Define $X_0 = 0$. Then

$$b\nu(0, b, n) \leq \sum_{j=1}^{n} \eta_j (X_j - X_{j-1}) = \sum_{j=1}^{n} \mathrm{I}(\eta_j = 1)(X_j - X_{j-1}).$$

The event

$$\{\eta_j = 1\} = \bigcup_k \{\tau_{2k-1} < j \leq \tau_{2k}\} = \bigcup_k \overbrace{\Big\{\tau_{2k-1} \leq j-1\Big\}}^{\in \mathcal{B}_{j-1}} \Big\backslash \overbrace{\Big\{\tau_{2k} \leq j-1\Big\}^c}^{\in \mathcal{B}_{j-1}} \in \mathcal{B}_{j-1}$$

i.e. the fact that at time j we are crossing upward is determined completely by the sequence up to time $j-1$. Then

$$\begin{aligned} b\mathbb{E}\nu(0, b, n) &\leq \sum_{j=1}^{n} \mathbb{E}\mathbb{E}\Big(\mathrm{I}(\eta_j = 1)(X_j - X_{j-1})\Big|\mathcal{B}_{j-1}\Big) = \sum_{j=1}^{n} \mathbb{E}\mathrm{I}(\eta_j = 1)\mathbb{E}(X_j - X_{j-1}|\mathcal{B}_{j-1}) \\ &= \sum_{j=1}^{n} \mathbb{E}\mathrm{I}(\eta_j = 1)(\mathbb{E}(X_j|\mathcal{B}_{j-1}) - X_{j-1}) \leq \sum_{j=1}^{n} \mathbb{E}(X_j - X_{j-1}) = \mathbb{E}X_n, \end{aligned}$$

where in the last inequality we used that $(X_j, \mathcal{B}_j)$ is a submartingale, $\mathbb{E}(X_j|\mathcal{B}_{j-1}) \geq X_{j-1}$, which implies that

$$\mathrm{I}(\eta_j = 1)(\mathbb{E}(X_j|\mathcal{B}_{j-1}) - X_{j-1}) \leq \mathbb{E}(X_j|\mathcal{B}_{j-1}) - X_{j-1}.$$

This finishes the proof.

□

Section 15

Convergence of martingales. Fundamental Wald's identity.

We finally get to our main result about the convergence of martingales and submartingales.

Theorem 34 *Let $(X_n, \mathcal{B}_n)_{-\infty<n<\infty}$ be a submartingale.*

1. *The limit $X_{-\infty} = \lim_{n\to-\infty} X_n$ exists a.s. and $\mathbb{E}X^+_{-\infty} < +\infty$. If $\mathcal{B}_{-\infty} = \bigcap_n \mathcal{B}_n$ then $(X_n, \mathcal{B}_n)_{-\infty\le n<\infty}$ is a submartingale.*
2. *If $\sup_n \mathbb{E}X_n^+ < \infty$, then $X_{+\infty} = \lim_{n\to\infty} X_n$ exists a.s. and $\mathbb{E}X^+_{+\infty} < \infty$.*
3. *If $(\max(X_n, a))_{-\infty<n<\infty}$ is uniformly integrable for any $a \in \mathbb{R}$ then $(X_n, \mathcal{B}_n)_{-\infty<n\le+\infty}$ is a submartingale where $X_{+\infty}$ is from part 2 and $\mathcal{B}_{+\infty} = \sigma\Big(\bigcup_n \mathcal{B}_n\Big)$.*

Remark. The first statement says that a *reverse* submartingale always converges (because it is right-closed by definition). The second statement says that under integrability conditions a submartingale converges. The last statement means that under stronger uniform intergrability conditions the submartingale not only converges but the limit right-closes the submartingale.

Proof. Let us note that X_n converges as $n \to \infty$ ($\infty = \pm\infty$) if and only if

$$\limsup X_n = \liminf X_n.$$

X_n diverges if and only if the following event occurs

$$\Big\{\limsup X_n > \liminf X_n\Big\} = \bigcup_{a<b} \Big\{\limsup X_n \ge b > a \ge \liminf X_n\Big\},$$

where the union is taken over all rational numbers $a < b$. In other words, $\mathbb{P}(X_n \text{ diverges}) > 0$ iff there exist rational $a < b$ such that

$$\mathbb{P}\Big(\limsup X_n \ge b > a \ge \liminf X_n\Big) > 0.$$

If we recall that $\nu(a, b, n)$ denotes the number of upcrossings of $[a, b]$, this is equivalent to

$$\mathbb{P}\Big(\lim_{n\to\infty} \nu(a, b, n) = \infty\Big) > 0.$$

Proof of 1. Let us define a new sequence

$$Y_1 = X_{-n}, Y_2 = X_{-n+1}, \ldots, Y_n = X_{-1}.$$

By Doob's inequality,

$$\mathbb{E}\nu_Y(a,b,n) \le \frac{\mathbb{E}(Y_n-a)^+}{b-a} = \frac{\mathbb{E}(X_{-1}-a)^+}{b-a} < \infty.$$

Since $0 \le \nu(a,b,n) \uparrow \nu(a,b)$, by monotone convergence theorem,

$$\mathbb{E}\nu(a,b) = \lim_{n\to\infty} \mathbb{E}\nu(a,b,n) < \infty.$$

Therefore, $\mathbb{P}(\nu(a,b)=\infty)=0$ which means that the limit

$$X_{-\infty} = \lim_{n\to-\infty} X_n$$

exists. By Fatou's lemma and submartingale property

$$\mathbb{E}X_{-\infty}^+ \le \liminf_{n\to-\infty} \mathbb{E}X_{-n}^+ \le \mathbb{E}X_{-1}^+ < \infty.$$

It remains to show that $(X_n,\mathcal{B}_n)_{-\infty\le n<\infty}$ is a submartingale. First of all, $X_{-\infty}=\lim_{n\to-\infty}X_n$ is measurable on $\mathcal{B}_m$ for all m and, therefore, measurable on $\mathcal{B}_{-\infty}=\cap\mathcal{B}_m$. Let us take a set $A\in\cap\mathcal{B}_n$. We would like to show that for any m

$$\mathbb{E}X_{-\infty}\mathrm{I}_A \le \mathbb{E}X_m\mathrm{I}_A.$$

Since $(X_n,\mathcal{B}_n)_{-\infty<n\le 0}$ is a right-closed submartingale, by Lemma 29, part 2,

$$\Big(\max(X_n,a),\mathcal{B}_n\Big)_{-\infty<n\le 0}$$

is uniformly integrable for any $a\in\mathbb{R}$. Since

$$\max(X_{-\infty},a) = \lim_{n\to-\infty}\max(X_n,a),$$

by Lemma 30,

$$\mathbb{E}\max(X_{-\infty},a)\mathrm{I}_A = \lim_{n\to-\infty}\mathbb{E}\max(X_n,a)\mathrm{I}_A \le \mathbb{E}\max(X_m,a)\mathrm{I}_A$$

for any m. Letting $a\to-\infty$, by monotone convergence theorem, $\mathbb{E}X_{-\infty}\mathrm{I}_A\le\mathbb{E}X_m\mathrm{I}_A$.

Proof of 2. If $\nu(a,b,n)$ is a number of upcrossings of $[a,b]$ by $X_1,\dots,X_n$ then by Doob's inequality

$$(b-a)\mathbb{E}\nu(a,b,n) \le \mathbb{E}(X_n-a)^+ \le |a| + \mathbb{E}X_n^+ \le K < \infty.$$

Therefore, $\mathbb{E}\nu(a,b)<\infty$ for $\nu(a,b)=\lim_{n\to\infty}\nu(a,b,n)$ and, as above, $X_{+\infty}=\lim_{n\to+\infty}X_n$ exists. Since all X_n are measurable on $\mathcal{B}_{+\infty}$ so is $X_{+\infty}$. Finally, by Fatou's lemma

$$\mathbb{E}X_{+\infty}^+ \le \liminf_{n\to\infty}\mathbb{E}X_n^+ < \infty.$$

Notice that another condition, $\sup_n\mathbb{E}|X_n|<\infty$, would similarly imply

$$\mathbb{E}|X_{+\infty}| \le \liminf_{n\to\infty}\mathbb{E}|X_n| < \infty.$$

Proof of 3. By 2, the limit $X_{+\infty}$ exists. We want to show that for any set $A\in\mathcal{B}_m$

$$\mathbb{E}X_m\mathrm{I}_A \le \mathbb{E}X_{+\infty}\mathrm{I}_A.$$

Let us denote $a\vee b=\max(a,b)$. Since $(X_n\vee a)$ is uniformly integrable and $X_n\vee a\to X_{+\infty}\vee a$, by Lemma 30,

$$\mathbb{E}(X_{+\infty}\vee a)\mathrm{I}_A = \lim_{n\to\infty}\mathbb{E}(X_n\vee a)\mathrm{I}_A.$$

Since a function $x\to x\vee a$ is convex and increasing, $(X_n\vee a)$ is also a submartingale and, thus, for $n>m$,

$$\mathbb{E}(X_m\vee a)\mathrm{I}_A \le \mathbb{E}(X_n\vee a)\mathrm{I}_A.$$

Therefore, $\mathbb{E}(X_m\vee a)\mathrm{I}_A\le\mathbb{E}(X_{+\infty}\vee a)\mathrm{I}_A$ and letting $a\to-\infty$, by monotone convergence theorem, we get that $\mathbb{E}X_m\mathrm{I}_A\le\mathbb{E}X_{+\infty}\mathrm{I}_A$.

□

As a corollary we get the following.

Corollary 2 *Martingale $(X_n, \mathcal{B}_n)$ is right-closable iff it is uniformly integrable.*

To prove this, apply case 3 above to (X_n) and $(-X_n)$ which are both submartingales.

Theorem 35 *(Levy's convergence) Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space and X be a real-valued random variable on it. Given a sequence of σ-algebras*

$$\mathcal{B}_1 \subseteq \ldots \subseteq \mathcal{B}_n \subseteq \ldots \subseteq \mathcal{B}_{+\infty} \subseteq \mathcal{B}$$

where $\mathcal{B}_{+\infty} = \sigma\Big(\bigcup_{1\le n<\infty} \mathcal{B}_n\Big)$, we have $X_n = \mathbb{E}(X|\mathcal{B}_n) \to \mathbb{E}(X|\mathcal{B}_{+\infty})$ a.s.

Proof. $(X_n, \mathcal{B}_n)_{1\le n<\infty}$ is a right-closable martingale since $X_n = \mathbb{E}(X|\mathcal{B}_n)$. Therefore, it is uniformly integrable and by previous theorem the limit $X_{+\infty} = \lim_{n\to\infty} \mathbb{E}(X|\mathcal{B}_n)$ exists. It remains to show that

$$X_{+\infty} = \mathbb{E}(X|\mathcal{B}_{+\infty}).$$

For a set A in algebra $\bigcup \mathcal{B}_n$, $A \in \mathcal{B}_m$ for some m and by Lemma 30

$$\mathbb{E}X_{+\infty}\mathrm{I}_A = \lim_{n\to\infty} \mathbb{E}X_n\mathrm{I}_A = \mathbb{E}X_m\mathrm{I}_A = \mathbb{E}(\mathbb{E}(X|\mathcal{B}_m)\mathrm{I}_A) = \mathbb{E}X\mathrm{I}_A.$$

By uniqueness in the Caratheodory extension theorem, $\mathbb{E}X_{+\infty}\mathrm{I}_A = \mathbb{E}X\mathrm{I}_A$ for all $A \in \sigma\Big(\bigcup_n \mathcal{B}_n\Big)$ which means that $X_{+\infty} = \mathbb{E}(X|\mathcal{B}_{+\infty})$.

□

Exercise. *(Kolmogorov's 0-1 law)* Consider arbitrary random variables $(X_i)_{i\ge1}$ and consider σ-algebras $\mathcal{B}_n = \sigma(X_1, \ldots, X_n)$ and $\mathcal{B}_\infty = \sigma((X_l)_{l\ge1})$. A set $A \in \mathcal{B}_\infty$ is a *tail event* if it is independent of $\mathcal{B}_n$ for any n. If we consider conditional expectations $\mathbb{E}(I_A|\mathcal{B}_n)$ then using Levy's convergence theorem and the fact that $\mathbb{E}(I_A|\mathcal{B}_n)$ and I_A are independent one can prove that $\mathbb{P}(A) = 0$ or 1. This gives a martingale proof of improved Kolmogorov's 0-1 law.

Examples.

1. *(Strong law of large numbers)* Let $(X_n)_{n\ge1}$ be i.i.d. Let

$$\mathcal{B}_{-n} = \sigma(S_n, S_{n+1}, \ldots), \;\; S_n = X_1 + \cdots + X_n.$$

We showed before that

$$Z_{-n} = \frac{S_n}{n} = \mathbb{E}(X_1|\mathcal{B}_{-n}),$$

i.e. $(Z_{-n}, \mathcal{B}_{-n})$ is a reverse martingale and, therefore, the limit $Y = \lim_{n\to+\infty} Z_{-n}$ exists. Y is measurable on σ-algebra $\mathcal{B}_{-\infty} = \bigcap_n \mathcal{B}_{-n}$. Since each event in $\mathcal{B}_{-\infty}$ is symmetric, i.e. invariant under permutations of X_n's, by Savage-Hewitt 0 - 1 law, the probability of each event is 0 or 1, i.e. $\mathcal{B}_{-\infty}$ consists of $\emptyset$ and Ω up to sets of measure zero. Therefore, Y is a constant a.s. and since $(Z_n)_{-\infty\le n\le-1}$ is also a martingale, $\mathbb{E}Y = \mathbb{E}X_1$. Therefore, $S_n/n \to \mathbb{E}X_1$ a.s.

2. *(Kolmogorov's law of large numbers)* Consider a sequence $(X_n)_{n\ge1}$ such that

$$\mathbb{E}(X_{n+1}|X_1, \ldots, X_n) = 0.$$

We do not assume independence. If a sequence (b_n) is such that

$$b_n < b_{n+1}, \;\; \lim_{n\to\infty} b_n = \infty \;\; \text{and} \;\; \sum_{n=1}^{\infty} \frac{\mathbb{E}X_n^2}{b_n^2} < \infty$$

then $S_n/n \to 0$ a.s. Indeed, $Y_n = \sum_{k\le n}(X_k/b_k)$ is a martingale and (Y_n) is uniformly integrable since

$$\mathbb{E}|Y_n|\mathrm{I}(Y_n| > M) \le \frac{1}{M}\mathbb{E}|Y_n|^2 \le \frac{1}{M}\sum_{n=1}^{\infty} \frac{\mathbb{E}X_k^2}{b_k^2} \to 0$$

as $M \to \infty$. Therefore, the limit $Y = \lim Y_n$ exists and it is an easy exercise to show that $S_n/b_n \to 0$ (Kronecker's lemma).

3. *(Polya' urn scheme)* Let us recall the Polya urn scheme from Section 5. Let us consider a sequence

$$Y_n = \frac{\#(\text{blue balls after } n \text{ iterations})}{\#(\text{total after } n \text{ iterations})}.$$

Y_n is a martingale because given that at step n the numbers of blue and red balls are b and r, the expected number of balls at step $n+1$ will be

$$\mathbb{E}(Y_{n+1}|\mathcal{B}_n) = \frac{b}{b+r} \cdot \frac{b+c}{b+r+c} + \frac{r}{b+r} \cdot \frac{b}{b+r+c} = \frac{b}{b+r} = Y_n.$$

Since Y_n is bounded, by martingale convergence theorem, the limit $Y = \lim_{n\to\infty} Y_n$ exists. What is the distribution of Y? Let us consider a sequence

$$X_i = \begin{cases} 1 & \text{blue at step } i \\ 0 & \text{red at step } i \end{cases}$$

and let $S_n = \sum_{i \le n} X_i$. Clearly,

$$Y_n = \frac{b + S_n c}{b + r + nc} \approx \frac{S_n}{n}$$

as $n \to \infty$ and, therefore, $S_n/n \to Y$. The sequence (X_n) is exchangeable and by de Finetti's theorem in Section 5 we showed that

$$\mathbb{P}(S_n = k) = \binom{n}{k} \int_0^1 x^k (1-x)^{n-k} d\beta\Big(\frac{b}{c}, \frac{r}{c}\Big)(x).$$

For any function $u \in C([0,1])$,

$$\mathbb{E}u\Big(\frac{S_n}{n}\Big) = \sum_{k=0}^n u\Big(\frac{k}{n}\Big)\binom{n}{k} \int_0^1 x^k (1-x)^{n-k} d\beta\Big(\frac{b}{c}, \frac{r}{c}\Big)(x) = \int_0^1 B_n(x) d\beta\Big(\frac{b}{c}, \frac{r}{c}\Big)(x),$$

where $B_n(x)$ is the Bernstein polynomial that approximates $u(x)$ uniformly on $[0,1]$. Therefore,

$$\mathbb{E}u\Big(\frac{S_n}{n}\Big) \to \int_0^1 u(x) d\beta\Big(\frac{b}{c}, \frac{r}{c}\Big)(x)$$

which means that

$$\mathcal{L}\Big(\frac{S_n}{n}\Big) \to \beta\Big(\frac{b}{c}, \frac{r}{c}\Big) = \mathcal{L}(Y),$$

i.e. the limit Y has Beta distribution $\beta\big(\frac{b}{c}, \frac{r}{c}\big)$.

□

Optional stopping for martingales revisited. Let τ be a stopping time. We would like to determine when $\mathbb{E}X_\tau = \mathbb{E}X_1$. As we saw above in the case of two stopping times, some kind of integrability assumptions are necessary. In this simpler case, the necessary conditions are clear from the proof.

Lemma 31 *We have*

$$\mathbb{E}X_1 = \lim_{n\to\infty} \mathbb{E}X_\tau \mathrm{I}(\tau \le n) \iff \lim_{n\to\infty} \mathbb{E}X_n \mathrm{I}(\tau \ge n) = 0.$$

Proof. We can write,

$$\begin{aligned} \mathbb{E}X_\tau \mathrm{I}(\tau \le n) = \sum_{1\le k\le n} \mathbb{E}X_k \mathrm{I}(\tau = k) &= \sum_{1\le k\le n} \Big(\mathbb{E}X_k \mathrm{I}(\tau \ge k) - \mathbb{E}X_k \mathrm{I}(\tau \ge k+1)\Big) \\ \Big\{\text{since } \{\tau \ge k+1\} = \{\tau \le k\}^c \in \mathcal{B}_k\Big\} &= \sum_{1\le k\le n} \Big(\mathbb{E}X_k \mathrm{I}(\tau \ge k) - \mathbb{E}X_{k+1} \mathrm{I}(\tau \ge k+1)\Big) \\ &= \mathbb{E}X_1 - \mathbb{E}X_{n+1}\mathrm{I}(\tau \ge n+1). \end{aligned}$$

Example. Given $0 < p < 1$, consider i.i.d. random variables $(\xi_i)_{i\geq 1}$ such that

$$\mathbb{P}(\xi_i = 1) = p, \;\; \mathbb{P}(\xi_i = -1) = 1 - p$$

and consider a random walk $X_0 = 0, X_{n+1} = X_n + \xi_{n+1}$. Consider two integers $a \leq -1$ and $b \geq 1$ and define a stopping time

$$\tau = \min\{k \geq 1, X_n = a \text{ or } b\}.$$

If $q = 1 - p$ then $Y_n = (q/p)^{X_n}$ is a martingale since

$$\mathbb{E}(Y_{n+1}|\mathcal{B}_n) = p\Big(\frac{q}{p}\Big)^{X_n+1} + q\Big(\frac{q}{p}\Big)^{X_n-1} = \Big(\frac{q}{p}\Big)^{X_n}.$$

It is easy to show that $\mathbb{P}(\tau \geq n) \to 0$ as $n \to \infty$ and using previous lemma,

$$1 = \mathbb{E}Y_0 = \mathbb{E}Y_\tau = \Big(\frac{q}{p}\Big)^b \mathbb{P}(X_\tau = b) + \Big(\frac{q}{p}\Big)^a (1 - \mathbb{P}(X_\tau = b)).$$

If $p \neq 1/2$ we can solve this

$$\mathbb{P}(X_\tau = b) = \frac{(q/p)^a - 1}{(q/p)^a - (q/p)^b}.$$

If $q = p = 1/2$ then X_n is a martingale and

$$0 = b\mathbb{P}(X_\tau = b) + a(1 - \mathbb{P}(X_\tau = b)) \Longrightarrow \mathbb{P}(X_\tau = b) = \frac{a}{a - b}.$$

Exercise. Compute $\mathbb{E}\tau$. *Hint:* for $p \neq 1/2$, use that $X_n - n\mathbb{E}\xi_1$ is a martingale; for $p = 1/2$, use that $X_n^2 - n$ is a martingale.

□

Fundamental Wald's identity. Let $(X_n)_{n\geq 1}$ be an i.i.d. sequence of random variables and suppose that a *Laplace transform* $\varphi(\lambda) = \mathbb{E}e^{\lambda X_1}$ is defined on some nontrivial interval (λ_-, λ_+) containing 0. Since

$$\mathbb{E}e^{\lambda S_n} = (\mathbb{E}e^{\lambda X_1})^n = \varphi(\lambda)^n \Longrightarrow \frac{e^{\lambda S_n}}{\varphi(\lambda)^n} \quad \text{is a martingale.}$$

Let τ be a stopping time. If

$$\mathbb{E}\frac{e^{\lambda S_n}}{\varphi(\lambda)^n}\mathrm{I}(\tau \geq n) \to 0 \tag{15.0.1}$$

then by the above lemma we get

$$\mathbb{E}\frac{e^{\lambda S_\tau}}{\varphi(\lambda)^\tau}\mathrm{I}(\tau < \infty) = \mathbb{E}\frac{e^{\lambda X_1}}{\varphi(\lambda)} = 1$$

which is called the *fundamental Wald's identity.* In some cases one can use this to obtain Laplace transform of a stopping time and, thus, its distribution.

Example. Let $X_0 = 0, \mathbb{P}(X_i = \pm 1) = 1/2, S_n = \sum_{k\leq n} X_k$. Given integer $z \geq 1$, let

$$\tau = \min\{k : S_k = -z \text{ or } z\}.$$

Since $\varphi(\lambda) = \mathrm{ch}(\lambda) \geq 1$,

$$\mathbb{E}\frac{e^{\lambda S_n}}{\varphi(\lambda)^n}\mathrm{I}(\tau \geq n) \leq \frac{e^{|\lambda| z}}{\mathrm{ch}(\lambda)^n}\mathbb{P}(\tau \geq n)$$

and (15.0.1) is satisfied. Therefore,

$$1 = \mathbb{E}\frac{e^{\lambda S_\tau}}{\varphi(\lambda)^\tau} = e^{\lambda z}\mathbb{E}\mathrm{ch}(\lambda)^{-\tau}\mathrm{I}(S_\tau = z) + e^{-\lambda z}\mathbb{E}\mathrm{ch}(\lambda)^{-\tau}\mathrm{I}(S_\tau = -z) = \frac{e^{\lambda z} + e^{-\lambda z}}{2}\mathbb{E}\mathrm{ch}(\lambda)^{-\tau}$$

by symmetry. Therefore,

$$\mathbb{E}\mathrm{ch}(\lambda)^{-\tau} = \frac{1}{\mathrm{ch}(\lambda z)} \quad \text{and} \quad \mathbb{E}e^{\gamma\tau} = \frac{1}{\mathrm{ch}(\mathrm{ch}^{-1}(e^{-\gamma})z)}$$

by a change of variables $e^{\gamma} = 1/\mathrm{ch}\lambda$.

□

For more general stopping times the condition (15.0.1) might not be easy to check. We will now show another approach that is helpful to verify a fundamental Wald's identity. If $\mathbb{P}$ is the distribution of X_i's, let $\mathbb{P}_\lambda$ be a distribution with Radon-Nikodym derivative w.r.t. $\mathbb{P}$ given by

$$\frac{d\mathbb{P}_\lambda}{d\mathbb{P}} = \frac{e^{\lambda x}}{\varphi(\lambda)}.$$

This is, indeed, a density since

$$\int_{\mathbb{R}} \frac{e^{\lambda x}}{\varphi(x)} d\mathbb{P} = \frac{\varphi(\lambda)}{\varphi(\lambda)} = 1.$$

We will think of (X_n) as defined on the product space $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mathbb{P}^\infty)$. For example, a set

$$\{\tau = n\} \in \sigma(X_1, \ldots, X_n) \subseteq \mathcal{B}^n$$

is a Borel set on $\mathbb{R}^n$. We can write,

$$\begin{aligned}
\mathbb{E}\frac{e^{\lambda S_\tau}}{\varphi(\lambda)^\tau}\mathrm{I}(\tau < \infty) = \sum_{n=1}^\infty \mathbb{E}\frac{e^{\lambda S_n}}{\varphi(\lambda)^n}\mathrm{I}(\tau = n) &= \sum_{n=1}^\infty \int\limits_{\{\tau=n\}} \frac{e^{\lambda(x_1+\cdots+x_n)}}{\varphi(\lambda)^n}\, d\mathbb{P}(x_1)\ldots d\mathbb{P}(x_n) \\
&= \sum_{n=1}^\infty \int\limits_{\{\tau=n\}} d\mathbb{P}_\lambda(x_1)\ldots d\mathbb{P}_\lambda(x_n) = \mathbb{P}_\lambda(\tau < \infty).
\end{aligned}$$

This means that we can think of (X_n) as having a distribution $\mathbb{P}_\lambda$ and to prove Wald's identity we need to show that $\mathbb{P}_\lambda(\tau < \infty) = 1$.

Example. *(Crossing a growing boundary)* Suppose that we have a boundary given by a sequence $(f(k))$ that changes with time and a stopping time (crossing time):

$$\tau = \min\{k : S_k \geq f(k)\}.$$

To verify (15.0.1), we need to show that $\mathbb{P}_\lambda(\tau < \infty) = 1$. Under the law $\mathbb{P}_\lambda$, random variables X_i have expectation

$$\mathbb{E}_\lambda X_i = \int x \frac{e^{\lambda x}}{\varphi(\lambda)} d\mathbb{P}(x) = \frac{\varphi'(\lambda)}{\varphi(\lambda)}.$$

By the strong law of large numbers

$$\lim_{n\to\infty} \frac{S_n}{n} = \frac{\varphi'(\lambda)}{\varphi(\lambda)}$$

and, therefore, if the growth of the crossing boundary satisfies

$$\limsup_{n\to\infty} \frac{f(n)}{n} < \frac{\varphi'(\lambda)}{\varphi(\lambda)}$$

then, obviously, $\mathbb{P}_\lambda(\tau < \infty) = 1$.

□

Section 16

Convergence on metric spaces. Portmanteau Theorem. Lipschitz Functions.

Let (S, d) be a metric space and $\mathcal{B}$ - a Borel σ-algebra generated by open sets. Let us recall that $\mathbb{P}_n \to \mathbb{P}$ weakly on $\mathcal{B}$ if

$$\int f d\mathbb{P}_n \to \int f d\mathbb{P}$$

for all $f \in C_b(S)$ - real-valued bounded continuous functions on S.

For a set $A \subseteq S$, we denote by $\bar{A}$ the closure of A, $\text{int}A$ - interior of A and $\partial A = \bar{A} \setminus \text{int}A$ - boundary of A. A is called a *continuity set* of $\mathbb{P}$ if $\mathbb{P}(\partial A) = 0$.

Theorem 36 *(Portmanteau theorem) The following are equivalent.*

1. $\mathbb{P}_n \to \mathbb{P}$ *weakly.*
2. *For any open set* $U \subseteq S$, $\liminf_{n\to\infty} \mathbb{P}_n(U) \geq \mathbb{P}(U)$.
3. *For any closed set* $F \subseteq S$, $\limsup_{n\to\infty} \mathbb{P}_n(F) \leq \mathbb{P}(F)$.
4. *For any continuity set* A *of* $\mathbb{P}$, $\lim_{n\to\infty} \mathbb{P}_n(A) = \mathbb{P}(A)$.

Proof.

1$\Longrightarrow$2. Let U be an open set and $F = U^c$. Consider a sequence of functions in $C_b(S)$

$$f_m(s) = \min(1, md(s, F))$$

such that $f_m(s) \uparrow \text{I}_U(s)$. (This is not necessarily true if U is not open.) Since $\mathbb{P}_n \to \mathbb{P}$,

$$\mathbb{P}_n(U) \geq \int f_m d\mathbb{P}_n \to \int f_m d\mathbb{P} \quad \text{as } n \to \infty \text{ and } \quad \liminf_{n\to\infty} \mathbb{P}_n(U) \geq \int f_m d\mathbb{P}.$$

Letting $m \to \infty$, by monotone convergence theorem.

$$\liminf_{n\to\infty} \mathbb{P}_n(U) \geq \int \text{I}_U d\mathbb{P} = \mathbb{P}(U).$$

2$\Longleftrightarrow$3. By taking complements.

2, 3$\Longrightarrow$4. Since $\text{int}A$ is open and $\bar{A}$ is closed and $\text{int}A \subseteq \bar{A}$, by 2 and 3,

$$\mathbb{P}(\text{int}A) \leq \liminf_{n\to\infty} \mathbb{P}_n(\text{int}A) \leq \limsup_{n\to\infty} \mathbb{P}_n(\bar{A}) \leq \mathbb{P}(\bar{A}).$$

If $\mathbb{P}(\partial A) = 0$ then $\mathbb{P}(\bar{A}) = \mathbb{P}(\text{int}A) = \mathbb{P}(A)$ and, therefore, $\lim \mathbb{P}_n(A) = \mathbb{P}(A)$.

4$\Longrightarrow$1. Consider $f \in C_b(S)$ and let $F_y = \{s \in S : f(s) = y\}$ be a level set of f. There exist at most countably many y such that $\mathbb{P}(F_y) > 0$. Therefore, for any $\varepsilon > 0$ we can find a sequence $a_1 \leq \ldots \leq a_N$ such that

$$\max(a_{k+1} - a_k) \leq \varepsilon, \;\; \mathbb{P}(F_{a_k}) = 0 \quad \text{for all } k$$

and the range of f is inside the interval (a_1, a_N). Let

$$B_k = \{s \in S : a_k \leq f(s) < a_{k+1}\} \quad \text{and} \quad f_\varepsilon(s) = \sum a_k \text{I}(s \in B_k).$$

Since f is continuous, $\partial B_k \subseteq F_{a_k} \cup F_{a_{k+1}}$ and $\mathbb{P}(\partial B_k) = 0$. By 4,

$$\int f_\varepsilon d\mathbb{P}_n = \sum_k a_k \mathbb{P}_n(B_k) \to \sum_k a_k \mathbb{P}(B_k) = \int f_\varepsilon d\mathbb{P}.$$

Since, by construction, $|f_\varepsilon(s) - f(s)| \leq \varepsilon$, letting $\varepsilon \to 0$ proves that $\int f d\mathbb{P}_n \to \int f d\mathbb{P}$.

□

Lipschitz functions. For a function $f : S \to \mathbb{R}$, let us define a Lipschitz semi-norm by

$$||f||_{\text{L}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)}.$$

Clearly, $||f||_{\text{L}} = 0$ iff f is constant so $||f||_{\text{L}}$ is not a norm. Let us define a *bounded Lipschitz* norm by

$$||f||_{\text{BL}} = ||f||_{\text{L}} + ||f||_\infty,$$

where $||f||_\infty = \sup_{s\in S} |f(s)|$. Let

$$BL(S, d) = \Big\{ f : S \to \mathbb{R} \, : \, ||f||_{\text{BL}} < \infty \Big\}$$

be a set of all bounded Lipschitz functions.

Lemma 32 *If $f, g \in BL(S, d)$ then $fg \in BL(S, d)$ and $||fg||_{\text{BL}} \leq ||f||_{\text{BL}} ||g||_{\text{BL}}$.*

Proof. First of all, $||fg||_\infty \leq ||f||_\infty ||g||_\infty$. We can write,

$$\begin{aligned} |f(x)g(x) - f(y)g(y)| &\leq |f(x)(g(x) - g(y))| + |g(y)(f(x) - f(y))| \\ &\leq ||f||_\infty ||g||_{\text{L}} d(x, y) + ||g||_\infty ||f||_{\text{L}} d(x, y) \end{aligned}$$

and, therefore,

$$||fg||_{\text{BL}} \leq ||f||_\infty ||g||_\infty + ||f||_\infty ||g||_{\text{L}} + ||g||_\infty ||f||_{\text{L}} \leq ||f||_{\text{BL}} ||g||_{\text{BL}}.$$

□

Let us recall the notations $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$ and let $* = \wedge$ or $\vee$.

Lemma 33 *The following hold.*

1. $||f_1 * \cdots * f_k||_{\text{L}} \leq \max_{1\leq i\leq k} ||f_i||_{\text{L}}$.
2. $||f_1 * \cdots * f_k||_{\text{BL}} \leq 2 \max_{1\leq i\leq k} ||f_i||_{\text{BL}}$.

Proof. *Proof of 1.* It is enough to consider $k = 2$. For specificity, take $* = \vee$. Given $x, y \in S$, suppose that

$$f_1 \vee f_2(x) \geq f_1 \vee f_2(y) = f_1(y).$$

Then

$$\begin{aligned} |f_1 \vee f_2(y) - f_1 \vee f_2(x)| = f_1 \vee f_2(x) - f_1 \vee f_2(y) &\leq \begin{cases} f_1(x) - f_1(y), & \text{if } f_1(x) \geq f_2(x) \\ f_2(x) - f_2(y), & \text{otherwise} \end{cases} \\ &\leq ||f_1||_{\mathrm{L}} \vee ||f_2||_{\mathrm{L}} d(x, y). \end{aligned}$$

This finishes the proof of 1.

Proof of 2. First of all, obviously,

$$||f_1 * \cdots * f_k||_\infty \leq \max_{1 \leq i \leq k} ||f_i||_\infty.$$

Therefore, using 1,

$$||f_1 * \cdots * f_k||_{\mathrm{BL}} \leq \max_i ||f_i||_\infty + \max_i ||f_i||_{\mathrm{L}} \leq 2 \max_i ||f_i||_{\mathrm{BL}}.$$

□

Theorem 37 *(Extension theorem) Given a set $A \subseteq S$ and a bounded Lipschitz function $f \in BL(A, d)$ on A, there exists an extension $h \in BL(S, d)$ such that*

$$f = h \quad on \quad A \quad and \quad ||h||_{\mathrm{BL}} = ||f||_{\mathrm{BL}}.$$

Proof. Let us first find an extension such that $||h||_{\mathrm{L}} = ||f||_{\mathrm{L}}$. We will start by extending f to one point $x \in S \setminus A$. The value $y = h(x)$ must satisfy

$$|y - f(s)| \leq \|f\|_{\mathrm{L}} d(x, s) \quad \text{for all} \quad s \in A$$

or, equivalently,

$$\inf_{s \in A} (f(s) + ||f||_{\mathrm{L}} d(x, s)) \geq y \geq \sup_{s \in A} (f(s) - ||f||_{\mathrm{L}} d(x, s)).$$

Such y exists iff for all $s_1, s_2 \in A$,

$$f(s_1) + ||f||_{\mathrm{L}} d(x, s_1) \geq f(s_2) - ||f||_{\mathrm{L}} d(x, s_2).$$

This inequality is satisfied because by triangle inequality

$$f(s_2) - f(s_1) \leq ||f||_{\mathrm{L}} d(s_1, s_2) \leq ||f||_{\mathrm{L}} (d(s_1, x) + d(s_2, x)).$$

It remains to apply Zorn's lemma to show that f can be extended to the entire S. Define order by inclusion:

$$f_1 \prec f_2 \text{ if } f_1 \text{ is defined on } A_1,\ f_2 \text{ - on } A_2,\ A_1 \subseteq A_2,\ f_1 = f_2 \text{ on } A_1 \text{ and } \|f_1\|_{\mathrm{L}} = \|f_2\|_{\mathrm{L}}.$$

For any chain $\{f_\alpha\}, f = \bigcup f_\alpha \succ f_\alpha$. By Zorn's lemma there exists a maximal element h. It is defined on the entire S because, otherwise, we could extend to one more point. To extend preserving BL norm take

$$h' = (h \wedge ||f||_\infty) \vee (-||f||_\infty).$$

By part 1 of previous lemma, it is easy to see that $||h'||_{\mathrm{BL}} = ||f||_{\mathrm{BL}}$.

□

Stone-Weierstrass Theorem.

A set $A \subseteq S$ is *totally bounded* if for any $\varepsilon > 0$ there exists a finite ε-cover of A, i.e. a set of points $a_1, \ldots, a_N$ such that

$$A \subseteq \bigcup_{i \leq N} B(a_i, \varepsilon),$$

where $B(a, \varepsilon) = \{y \in S : d(a, y) \leq \varepsilon\}$ is a ball of radius ε centered at a. Let us recall the following theorem from analysis.

Theorem 38 *(Arzela-Ascoli) Let (S,d) be a compact metric space and let $(C(S),d_\infty)$ be the space of continuous real-valued functions on S with uniform convergence metric*

$$d_\infty(f,g)=\sup_{x\in S}|f(x)-g(x)|.$$

A subset $\mathcal{F}\subseteq C(S)$ is totally bounded in d_∞ metric iff $\mathcal{F}$ is equicontinuous and uniformly bounded.

Remark. Equicontinuous means that for any $\varepsilon >$ there exists $\delta>0$ such that if $d(x,y)\le\delta$ then for all $f\in\mathcal{F}$, $|f(x)-f(y)|\le\varepsilon$.

Theorem 39 *(Stone-Weierstrass) Let (S,d) be a compact metric space and $\mathcal{F}\subseteq C(S)$ is such that*

1. *$\mathcal{F}$ is algebra, i.e. for all $f,g\in\mathcal{F}, c\in\mathbb{R}$, we have $cf+g\in\mathcal{F}, fg\in\mathcal{F}$.*
2. *$\mathcal{F}$ separates points, i.e. if $x\neq y\in S$ then there exists $f\in\mathcal{F}$ such that $f(x)\neq f(y)$.*
3. *$\mathcal{F}$ contains constants.*

Then $\mathcal{F}$ is dense in $C(S)$.

Corollary 3 *If (S,d) is a compact space then $BL(S,d)$ is dense in $C(S)$.*

Proof. For $\mathcal{F}=BL(S,d)$ in the Stone-Weierstrass theorem, 3 is obvious, 1 follows from Lemma 32 and 2 follows from the extension Theorem 37, since a function defined on two points $x\neq y$ such that $f(x)\neq f(y)$ can be extended to the entire S.

□

Proof of Theorem 39. Consider bounded $f\in\mathcal{F}$, i.e. $|f(x)|\le M$. A function $x\to|x|$ defined on the interval $[-M,M]$ can be uniformly approximated by polynomials of x by the Weierstrass theorem on the real line or, for example, using Bernstein's polynomials. Therefore, $|f(x)|$ can be uniformly approximated by polynomials of $f(x)$, and by properties 1 and 3, by functions in $\mathcal{F}$. Therefore, if $\bar{\mathcal{F}}$ is the closure of $\mathcal{F}$ in d_∞ norm then for any $f\in\bar{\mathcal{F}}$ its absolute value $|f|\in\bar{\mathcal{F}}$. Therefore, for any $f,g\in\bar{\mathcal{F}}$ we have

$$\min(f,g)=\frac{1}{2}(f+g)-\frac{1}{2}|f-g|\in\bar{\mathcal{F}},\ \ \max(f,g)=\frac{1}{2}(f+g)+\frac{1}{2}|f-g|\in\bar{\mathcal{F}}. \tag{16.0.1}$$

Given any points $x\neq y$ and $c,d\in\mathbb{R}$ one can always find $f\in\mathcal{F}$ such that $f(x)=c$ and $f(y)=d$. Indeed, by property 2 we can find $g\in\mathcal{F}$ such that $g(x)\neq g(y)$ and, as a result, a system of equations

$$ag(x)+b=c,\ \ ag(y)+b=d$$

has a solution a,b. Then the function $f=ag+b$ satisfies the above and it is in $\mathcal{F}$ by 1.

Take $h\in C(S)$ and fix x. For any y let $f_y\in\mathcal{F}$ be such that

$$f_y(x)=h(x),\ \ f_y(y)=h(y).$$

By continuity of f_y, for any $y\in S$ there exists an open neighborhood U_y of y such that

$$f_y(s)\ge h(s)-\varepsilon \text{ for } s\in U_y.$$

Since (U_y) is an open cover of the compact S, there exists a finite subcover $U_{y_1},\dots,U_{y_N}$. Let us define a function

$$f^x(s)=\max(f_{y_1}(s),\dots,f_{y_N}(s))\in\bar{\mathcal{F}} \text{ by (16.0.1)}.$$

By construction, it has the following properties:

$$f^x(x)=h(x),\ \ f^x(s)\ge h(s)-\varepsilon \quad\text{for all}\quad s\in S.$$

Again, by continuity of $f^x(s)$ there exists an open neighborhood U_x of x such that

$$f^x(s) \leq h(s) + \varepsilon \quad \text{for} \quad s \in U_x.$$

Take a finite subcover $U_{x_1}, \ldots, U_{x_M}$ and define

$$h'(s) = \min\bigl(f^{x_1}(s), \ldots, f^{x_M}(s)\bigr) \in \bar{\mathcal{F}} \text{ by (16.0.1).}$$

By construction, $h'(s) \leq h(s) + \varepsilon$ and $h'(s) \geq h(s) - \varepsilon$ for all $s \in S$ which means that $d_\infty(h', h) \leq \varepsilon$. Since $h' \in \bar{\mathcal{F}}$, this proves that $\bar{\mathcal{F}}$ is dense in $C(S)$.

□

Corollary 4 *If (S, d) is a compact space then $C(S)$ is separable in d_∞.*

Remark. Recall that this fact was used in the proof of the Selection Theorem, which was proved for general metric spaces.

Proof. By the above theorem, $BL(S, d)$ is dense in $C(S)$. For any integer $n \geq 1$, the set $\{f : ||f||_{\text{BL}} \leq n\}$ is uniformly bounded and equicontinuous. By the Arzela-Ascoli theorem, it is totally bounded and, therefore, separable which can be seen by taking finite $1/m$-covers for all $m \geq 1$. The union

$$\bigcup\{||f||_{\text{BL}} \leq n\} = BL(S, d)$$

is therefore separable in $C(S)$ which is, as a result, also separable.

□

Section 17

Metrics for convergence of laws. Empirical measures.

Levy-Prohorov metric. Consider a metric space (S, d). For a set $A \subseteq S$ let us denote by

$$A^\varepsilon = \{y \in S : d(x, y) < \varepsilon \text{ for some } x \in A\}$$

its ε-neighborhood. Let $\mathcal{B}$ be a Borel σ-algebra on S.

Definition. If $\mathbb{P}, \mathbb{Q}$ are probability distributions on $\mathcal{B}$ then

$$\rho(\mathbb{P}, \mathbb{Q}) = \inf\{\varepsilon > 0 : \mathbb{P}(A) \le \mathbb{Q}(A^\varepsilon) + \varepsilon \text{ for all } A \in \mathcal{B}\}$$

is called the *Levy-Prohorov distance* between $\mathbb{P}$ and $\mathbb{Q}$.

Lemma 34 *ρ is a metric on the set of probability laws on $\mathcal{B}$.*

Proof. 1. First, let us show that $\rho(\mathbb{Q}, \mathbb{P}) = \rho(\mathbb{P}, \mathbb{Q})$. Suppose that $\rho(\mathbb{P}, \mathbb{Q}) > \varepsilon$. Then there exists a set A such that $\mathbb{P}(A) > \mathbb{Q}(A^\varepsilon) + \varepsilon$. Taking complements gives

$$\mathbb{Q}(A^{\varepsilon c}) > \mathbb{P}(A^c) + \varepsilon \ge \mathbb{P}(A^{\varepsilon c \varepsilon}) + \varepsilon,$$

where the last inequality follows from the fact that $A^c \supseteq A^{\varepsilon c \varepsilon}$:

$$\begin{aligned} a \in A^{\varepsilon c \varepsilon} \Longrightarrow d(a, A^{\varepsilon c}) < \varepsilon \quad &\Longrightarrow \quad d(a, b) < \varepsilon \text{ for some } b \in A^{\varepsilon c} \\ &\qquad \Big\{\text{since } b \notin A^\varepsilon, d(b, A) \ge \varepsilon\Big\} \\ &\Longrightarrow \quad d(a, A) > 0 \Longrightarrow a \notin A \Longrightarrow a \in A^c. \end{aligned}$$

Therefore, for a set $B = A^{\varepsilon c}$, $\mathbb{Q}(B) > \mathbb{P}(B^\varepsilon) + \varepsilon$. This means that $\rho(\mathbb{Q}, \mathbb{P}) > \varepsilon$ and, therefore, $\rho(\mathbb{Q}, \mathbb{P}) \ge \rho(\mathbb{P}, \mathbb{Q})$. By symmetry, $\rho(\mathbb{Q}, \mathbb{P}) \le \rho(\mathbb{P}, \mathbb{Q})$ and $\rho(\mathbb{Q}, \mathbb{P}) = \rho(\mathbb{P}, \mathbb{Q})$.

2. Next, let us show that if $\rho(\mathbb{P}, \mathbb{Q}) = 0$ then $\mathbb{P} = \mathbb{Q}$. For any set F and any $n \ge 1$,

$$\mathbb{P}(F) \le \mathbb{Q}(F^{\frac{1}{n}}) + \frac{1}{n}.$$

If F is closed then $F^{\frac{1}{n}} \downarrow F$ as $n \to \infty$ and by continuity of measure

$$\mathbb{P}(F) \le \mathbb{Q}\Big(\bigcap F^{\frac{1}{n}}\Big) = \mathbb{Q}(F).$$

Similarly, $\mathbb{P}(F) \ge \mathbb{Q}(F)$ and, therefore, $\mathbb{P}(F) = \mathbb{Q}(F)$.

3. Finally, let us prove the triangle inequality

$$\rho(\mathbb{P},\mathbb{R}) \le \rho(\mathbb{P},\mathbb{Q}) + \rho(\mathbb{Q},\mathbb{R}).$$

If $\rho(\mathbb{P},\mathbb{Q}) < x$ and $\rho(\mathbb{Q},\mathbb{R}) < y$ then for any set A,

$$\mathbb{P}(A) \le \mathbb{Q}(A^x) + x \le \mathbb{R}\big((A^x)^y\big) + y + x \le \mathbb{R}\big(A^{x+y}\big) + x + y,$$

which means that $\rho(\mathbb{P},\mathbb{R}) \le x + y$.

□

Bounded Lipschitz metric. Given probability distributions $\mathbb{P},\mathbb{Q}$ on the metric space (S,d) we define a *bounded Lipschitz* distance between them by

$$\beta(\mathbb{P},\mathbb{Q}) = \sup\Big\{\Big|\int f d\mathbb{P} - \int f d\mathbb{Q}\Big| : ||f||_{\rm BL} \le 1\Big\}.$$

Lemma 35 *β is a metric on the set of probability laws on $\mathcal{B}$.*

Proof. $\beta(\mathbb{P},\mathbb{Q}) = \beta(\mathbb{Q},\mathbb{P})$ and the triangle inequality are obvious. It remains to prove that $\beta(\mathbb{P},\mathbb{Q}) = 0$ implies $\mathbb{P} = \mathbb{Q}$. Given a closed set F, the sequence of functions $f_m(x) = md(x,F) \wedge 1$ converges $f_m \uparrow {\rm I}_U$, where $U = F^c$. Obviously, $||f_m||_{\rm BL} \le m+1$ and, therefore, $\int f_m d\mathbb{P} = \int f_m d\mathbb{Q}$. Letting $m \to \infty$ proves that $\mathbb{P}(U) = \mathbb{Q}(U)$.

□

The law $\mathbb{P}$ on (S,d) is *tight* if for any $\varepsilon > 0$ there exists a compact $K \subseteq S$ such that $\mathbb{P}(S \setminus K) \le \varepsilon$.

Theorem 40 *(Ulam) If (S,d) is separable then for any law $\mathbb{P}$ on $\mathcal{B}$ there exists a closed totally bounded set $K \subseteq S$ such that $\mathbb{P}(S \setminus K) \le \varepsilon$. If (S,d) is complete and separable then K is compact and, therefore, every law is tight.*

Proof. Consider a sequence $\{x_1, x_2, \ldots\}$ that is dense in S. For any $m \ge 1$, $S = \bigcup_{i=1}^{\infty} \bar{B}\Big(x_i, \frac{1}{m}\Big)$, where $\bar{B}$ denotes a closed ball, and by continuity of measure, for large enough $n(m)$,

$$\mathbb{P}\Big(S \setminus \bigcup_{i=1}^{n(m)} \bar{B}\Big(x_i, \frac{1}{m}\Big)\Big) \le \frac{\varepsilon}{2^m}.$$

If we take

$$K = \bigcap_{m\ge 1} \bigcup_{i=1}^{n(m)} \bar{B}\Big(x_i, \frac{1}{m}\Big)$$

then

$$\mathbb{P}(S \setminus K) \le \sum_{m\ge 1} \frac{\varepsilon}{2^m} = \varepsilon.$$

K is closed and totally bounded by construction. If S is complete, K is compact.

□

Theorem 41 *Suppose that either (S,d) is separable or $\mathbb{P}$ is tight. Then the following are equivalent.*

1. $\mathbb{P}_n \to \mathbb{P}$.
2. *For all $f \in BL(S,d)$, $\int f d\mathbb{P}_n \to \int f d\mathbb{P}$.*
3. $\beta(\mathbb{P}_n,\mathbb{P}) \to 0$.
4. $\rho(\mathbb{P}_n,\mathbb{P}) \to 0$.

Proof. 1$\Longrightarrow$2. Obvious.

3$\Longrightarrow$4. In fact, we will prove that

$$\rho(\mathbb{P}_n, \mathbb{P}) \leq 2\sqrt{\beta(\mathbb{P}_n, \mathbb{P})}. \tag{17.0.1}$$

Given a Borel set $A \subseteq S$, consider a function

$$f(x) = 0 \vee \Big(1 - \frac{1}{\varepsilon} d(x, A)\Big) \quad \text{such that} \quad \mathrm{I}_A \leq f \leq \mathrm{I}_{A^\varepsilon}.$$

Obviously, $||f||_{\mathrm{BL}} \leq 1 + \varepsilon^{-1}$ and we can write

$$\begin{aligned} \mathbb{P}_n(A) &\leq \int f d\mathbb{P}_n = \int f d\mathbb{P} + \Big(\int f d\mathbb{P}_n - \int f d\mathbb{P}\Big) \\ &\leq \mathbb{P}(A^\varepsilon) + (1+\varepsilon^{-1}) \sup\Big\{\Big|\int f d\mathbb{P}_n - \int f d\mathbb{P}\Big| : ||f||_{\mathrm{BL}} \leq 1\Big\} \\ &= \mathbb{P}(A^\varepsilon) + (1+\varepsilon^{-1})\beta(\mathbb{P}_n, \mathbb{P}) \leq \mathbb{P}(A^\delta) + \delta, \end{aligned}$$

where $\delta = \max(\varepsilon, (1+\varepsilon^{-1})\beta(\mathbb{P}_n, \mathbb{P}))$. This implies that $\rho(\mathbb{P}_n, \mathbb{P}) \leq \delta$. Since ε is arbitrary we can minimize $\delta = \delta(\varepsilon)$ over ε. If we take $\varepsilon = \sqrt{\beta}$ then $\delta = \max(\sqrt{\beta}, \beta + \sqrt{\beta}) = \beta + \sqrt{\beta}$ and

$$\beta \leq 1 \Longrightarrow \rho \leq 2\sqrt{\beta}; \quad \beta \geq 1 \Longrightarrow \rho \leq 1 \leq 2\sqrt{\beta}.$$

4$\Longrightarrow$1. Suppose that $\rho(\mathbb{P}_n, \mathbb{P}) \to 0$ which means that there exists a sequence $\varepsilon_n \downarrow 0$ such that

$$\mathbb{P}_n(A) \leq \mathbb{P}(A^{\varepsilon_n}) + \varepsilon_n \quad \text{for all measurable } A \subseteq S.$$

If A is closed, then $\bigcap_{n\geq 1} A^{\varepsilon_n} = A$ and, by continuity of measure,

$$\limsup_{n\to\infty} \mathbb{P}_n(A) \leq \limsup_{n\to\infty} \Big(\mathbb{P}(A^{\varepsilon_n}) + \varepsilon_n\Big) = \mathbb{P}(A).$$

By the portmanteau theorem, $\mathbb{P}_n \to \mathbb{P}$.

2$\Longrightarrow$3. If $\mathbb{P}$ is tight, let K be a compact such that $\mathbb{P}(S \setminus K) \leq \varepsilon$. If (S, d) is separable, by Ulam's theorem, let K be a closed totally bounded set such that $\mathbb{P}(S \setminus K) \leq \varepsilon$. If we consider a function

$$f(x) = 0 \vee \Big(1 - \frac{1}{\varepsilon} d(x, K)\Big) \text{ with } ||f||_{\mathrm{BL}} \leq 1 + \frac{1}{\varepsilon}$$

then

$$\mathbb{P}_n(K^\varepsilon) \geq \int f d\mathbb{P}_n \to \int f d\mathbb{P} \geq \mathbb{P}(K) \geq 1 - \varepsilon,$$

which implies that for n large enough, $\mathbb{P}_n(K^\varepsilon) \geq 1 - 2\varepsilon$. This means that all $\mathbb{P}_n$ are essentially concentrated on K^ε. Let

$$B = \Big\{f : ||f||_{\mathrm{BL}(S,d)} \leq 1\Big\}, \;\; B_K = \Big\{f\big|_K : f \in B\Big\} \subseteq C(K),$$

where $f\big|_K$ denotes the restriction of f to K. If K is compact then, by the Arzela-Ascoli theorem, B_K is totally bounded with respect to d_∞. If K is totally bounded then we can isometrically identify functions in B_K with their unique extensions to the completion K' of K and, by the Arzela-Ascoli theorem for the compact K', B_K is again totally bounded with respect to d_∞. In any case, given $\varepsilon > 0$, we can find $f_1, \ldots, f_k \in B$ such that for all $f \in B$

$$\sup_{x\in K} |f(x) - f_j(x)| \leq \varepsilon \quad \text{for some } j \leq k.$$

This uniform approximation can also be extended to K^ε. Namely, for any $x \in K^\varepsilon$ take $y \in K$ such that $d(x, y) \leq \varepsilon$. Then

$$\begin{aligned} |f(x) - f_j(x)| &\leq |f(x) - f(y)| + |f(y) - f_j(y)| + |f_j(y) - f_j(x)| \\ &\leq ||f||_{\mathrm{L}} d(x, y) + \varepsilon + ||f_j||_{\mathrm{L}} d(x, y) \leq 3\varepsilon. \end{aligned}$$

Therefore, for any $f \in B$,

$$\begin{aligned}
\Big|\int f d\mathbb{P}_n - \int f d\mathbb{P}\Big| &\leq \Big|\int_{K^\varepsilon} f d\mathbb{P}_n - \int_{K^\varepsilon} f d\mathbb{P}\Big| + ||f||_\infty \big(\mathbb{P}_n(K^{\varepsilon c}) + \mathbb{P}(K^{\varepsilon c})\big) \\
&\leq \Big|\int_{K^\varepsilon} f d\mathbb{P}_n - \int_{K^\varepsilon} f d\mathbb{P}\Big| + 2\varepsilon + \varepsilon \\
&\leq \Big|\int_{K^\varepsilon} f_j d\mathbb{P}_n - \int_{K^\varepsilon} f_j d\mathbb{P}\Big| + 3\varepsilon + 3\varepsilon + 2\varepsilon + \varepsilon \\
&\leq \Big|\int f_j d\mathbb{P}_n - \int f_j d\mathbb{P}\Big| + 3\varepsilon + 3\varepsilon + 3\varepsilon + 2\varepsilon + \varepsilon \\
&\leq \max_{1\leq j\leq k}\Big|\int f_j d\mathbb{P}_n - \int f_j d\mathbb{P}\Big| + 12\varepsilon.
\end{aligned}$$

Finally,

$$\beta(\mathbb{P}_n, \mathbb{P}) = \sup_{f\in B}\Big|\int f d\mathbb{P}_n - \int f d\mathbb{P}\Big| \leq \max_{1\leq j\leq k}\Big|\int f_j d\mathbb{P}_n - \int f_j d\mathbb{P}\Big| + 12\varepsilon$$

and, using assumption 2, $\limsup_{n\to\infty} \beta(\mathbb{P}_n, \mathbb{P}) \leq 12\varepsilon$. Letting $\varepsilon \to 0$ finishes the proof.

□

Convergence of empirical measures. Let $(\Omega, \mathbb{P})$ be a probability space and $X_1, X_2, \ldots : \Omega \to S$ be an i.i.d. sequence of random variables with values in a metric space (S, d). Let μ be the law of X_i on S. Let us define the random *empirical measures* μ_n on the Borel σ-algebra $\mathcal{B}$ on S by

$$\mu_n(A)(\omega) = \frac{1}{n}\sum_{i=1}^n I(X_i(\omega) \in A), \;\; A \in \mathcal{B}.$$

By the strong law of large numbers, for any $f \in C_b(S)$,

$$\int f d\mu_n = \frac{1}{n}\sum_{i=1}^n f(X_i) \to \mathbb{E}f(X_1) = \int f d\mu \text{ a.s.}$$

However, the set of measure zero where this convergence is violated depends on f and it is not obvious that the convergence holds for all $f \in C_b(S)$ with probability one.

Theorem 42 *(Varadarajan) Let (S, d) be a separable metric space. Then μ_n converges to μ weakly almost surely,*

$$\mathbb{P}\big(\omega : \mu_n(\cdot)(\omega) \to \mu \text{ weakly}\big) = 1.$$

Proof. Since (S, d) is separable, by Theorem 2.8.2 in R.A.P., there exists a metric e on S such that (S, e) is totally bounded and e and d define the same topology, i.e. $e(s_n, s) \to 0$ if and only if $d(s_n, s) \to 0$. This, of course, means that $C_b(S, d) = C_b(S, e)$ and weak convergence of measures does not change. If (T, e) is the completion of (S, e) then (T, e) is compact. By the Arzela-Ascoli theorem, $BL(T, e)$ is separable with respect to the d_∞ norm and, therefore, $BL(S, e)$ is also separable. Let (f_m) be a dense subset of $BL(S, e)$. Then, by the strong law of large number,

$$\int f_m d\mu_n = \frac{1}{n}\sum_{i=1}^n f_m(X_i) \to \mathbb{E}f_m(X_1) = \int f_m d\mu \text{ a.s.}$$

Therefore, on the set of probability one, $\int f_m d\mu_n \to \int f_m d\mu$ for all $m \geq 1$. Since (f_m) is dense in $BL(S, e)$, on the same set of probability one, $\int f d\mu_n \to \int f d\mu$ for all $f \in BL(S, e)$. Since (S, e) is separable, the previous theorem implies that $\mu_n \to \mu$ weakly.

□

Section 18

Convergence and uniform tightness.

In this section, we will make several connections between convergence of measures and uniform tightness on general metric spaces, which are similar to the results in the Euclidean setting. First, we will show that, in some sense, uniform tightness is necessary for convergence of laws.

Theorem 43 *If $\mathbb{P}_n \to \mathbb{P}_0$ on S and each $\mathbb{P}_n$ is tight for $n \geq 0$, then $(\mathbb{P}_n)_{n\geq 0}$ is uniformly tight.*

Proof. Since $\mathbb{P}_n \to \mathbb{P}_0$ and $\mathbb{P}_0$ is tight, by Theorem 41, the Levy-Prohorov metric $\rho(\mathbb{P}_n, \mathbb{P}_0) \to 0$. Given $\varepsilon > 0$, let us take a compact K such that $\mathbb{P}_0(K) > 1 - \varepsilon$. By definition of ρ,

$$1 - \varepsilon < \mathbb{P}_0(K) \leq \mathbb{P}_n\Big(K^{\rho(\mathbb{P}_n,\mathbb{P}_0)+\frac{1}{n}}\Big) + \rho(\mathbb{P}_n, \mathbb{P}_0) + \frac{1}{n}$$

and, therefore,

$$a(n) = \inf\Big\{\delta > 0 : \mathbb{P}_n(K^\delta) > 1 - \varepsilon\Big\} \to 0.$$

By regularity of measure $\mathbb{P}_n$, any measurable set A can be approximated by its closed subset F. Since $\mathbb{P}_n$ is tight, we can choose a compact of measure close to one, and intersecting it with the closed subset F, we can approximate any set A by its compact subset. Therefore, there exists a compact $K_n \subseteq K^{2a(n)}$ such that $\mathbb{P}_n(K_n) > 1 - \varepsilon$. Let

$$L = K \bigcup (\cup_{n\geq 1} K_n).$$

Then $\mathbb{P}_n(L) \geq \mathbb{P}_n(K_n) > 1 - \varepsilon$. It remains to show that L is compact. Consider a sequence (x_n) on L. There are two possibilities. First, if there exists an infinite subsequence $(x_{n(k)})$ that belongs to one of the compacts K_j then it has a converging subsubsequence in K_j and as a result in L. If not, then there exists a subsequence $(x_{n(k)})$ such that $x_{n(k)} \in K_{m(k)}$ and $m(k) \to \infty$ as $k \to \infty$. Since

$$K_{m(k)} \subseteq K^{2a(m(k))}$$

there exists $y_k \in K$ such that

$$d(x_{n(k)}, y_k) \leq 2a(m(k)).$$

Since K is compact, the sequence $y_k \in K$ has a converging subsequence $y_{k(r)} \to y \in K$ which implies that $d(x_{n(k(r))}, y) \to 0$, i.e. $x_{n(k(r))} \to y \in L$. Therefore, L is compact.

□

We already know from the Selection Theorem in Section 8 that any uniformly tight sequence of laws on any metric space has a converging subsequence. Under additional assumptions on (S, d) we can complement the Selection Theorem and make some connections to the metrics defined in the previous section.

Theorem 44 *Let (S, d) be a complete separable metric space and A be a subset of probability laws on S. Then the following are equivalent.*

1. *A is uniformly tight.*

2. *For any sequence $\mathbb{P}_n \in A$ there exists a converging subsequence $\mathbb{P}_{n(k)} \to \mathbb{P}$ where $\mathbb{P}$ is a law on S.*

3. *A has the compact closure on the space of probability laws equipped with the Levy-Prohorov or bounded Lipschitz metrics ρ or β.*

4. *A is totally bounded with respect to ρ or β.*

Remark. Implications $1 \Longrightarrow 2 \Longrightarrow 3 \Longrightarrow 4$ hold without completeness assumption and the only implication where completeness will be used is $4 \Longrightarrow 1$.

Proof. $1 \Longrightarrow 2$. Any sequence $\mathbb{P}_n \in A$ is uniformly tight and, by selection theorem, there exists a converging subsequence.

$2 \Longrightarrow 3$. Since (S, d) is separable, by Theorem 41, $\mathbb{P}_n \to \mathbb{P}$ if and only if $\rho(\mathbb{P}_n, \mathbb{P})$ or $\beta(\mathbb{P}_n, \mathbb{P}) \to 0$. Every sequence in the closure $\bar{A}$ can be approximated by a sequence in A. That sequence has a converging subsequence that, obviously, converges to an element in $\bar{A}$ which means that the closure of A is compact.

$3 \Longrightarrow 4$. Compact sets are totally bounded and, therefore, if the closure $\bar{A}$ is compact, the set A is totally bounded.

$4 \Longrightarrow 1$. Since $\rho \leq 2\sqrt{\beta}$, we will only deal with ρ. For any $\varepsilon > 0$, there exists a finite subset $B \subseteq A$ such that $A \subseteq B^{\varepsilon}$. Since (S, d) is complete and separable, by Ulam's theorem, for each $\mathbb{P} \in B$ there exists a compact $K_{\mathbb{P}}$ such that $\mathbb{P}(K_{\mathbb{P}}) > 1 - \varepsilon$. Therefore,

$$K_B = \bigcup_{\mathbb{P} \in B} K_{\mathbb{P}} \text{ is a compact and } \mathbb{P}(K_B) > 1 - \varepsilon \text{ for all } \mathbb{P} \in B.$$

For any $\varepsilon > 0$, let F be a finite set such that $K_B \subseteq F^{\varepsilon}$ (here we will denote by F^{ε} the closed ε-neighborhood of F). Since $A \subseteq B^{\varepsilon}$, for any $\mathbb{Q} \in A$ there exists $\mathbb{P} \in B$ such that $\rho(\mathbb{Q}, \mathbb{P}) < \varepsilon$ and, therefore,

$$1 - \varepsilon \leq \mathbb{P}(K_B) \leq \mathbb{P}(F^{\varepsilon}) \leq \mathbb{Q}(F^{2\varepsilon}) + \varepsilon.$$

Thus, $1 - 2\varepsilon \leq \mathbb{Q}(F^{2\varepsilon})$ for all $\mathbb{Q} \in A$. Given $\delta > 0$, take $\varepsilon_m = \delta / 2^{m+1}$ and find F_m as above, i.e.

$$1 - \frac{\delta}{2^m} \leq \mathbb{Q}\Big(F_m^{\delta/2^m}\Big).$$

Then $\mathbb{Q}\Big(\bigcap_{m \geq 1} F_m^{\delta/2^m}\Big) \geq 1 - \sum_{m \geq 1} \frac{\delta}{2^m} = 1 - \delta$. Finally, $L = \bigcap_{m \geq 1} F_m^{\delta/2^m}$ is compact because it is closed and totally bounded by construction, and S is complete.

□

Corollary 5 *(Prohorov) The set of laws on a complete separable metric space is complete with respect to metrics ρ or β.*

Proof. If a sequence of laws is Cauchy w.r.t. ρ or β then it is totally bounded and by previous theorem it has a converging subsequence. Obviously, Cauchy sequence will converge to the same limit.

□

Finally, let us state as a result the idea which appeared in Lemma 19 in Section 9.

Lemma 36 *Suppose that $(\mathbb{P}_n)$ is uniformly tight on a metric space (S, d). Suppose that all converging subsequences $(\mathbb{P}_{n(k)})$ converge to the same limit, i.e. if $\mathbb{P}_{n(k)} \to \mathbb{P}_0$ then $\mathbb{P}_0$ is independent of $(n(k))$. Then $\mathbb{P}_n \to \mathbb{P}_0$.*

Proof. Any subsequence $(\mathbb{P}_{n(k)})$ is uniformly tight and, by the selection theorem, it has a converging subsubsequence $(\mathbb{P}_{n(k(r))})$ which has to converge to $\mathbb{P}_0$. Lemma 13 in Section 8 finishes the proof.

□

This will be very useful when proving convergence of laws on metric spaces, such as $C([0, 1])$, for example. If we can prove that $(\mathbb{P}_n)$ is uniformly tight and, assuming that a subsequence converges, can identify the unique limit, then the sequence $\mathbb{P}_n$ must converge to the same limit.

Section 19

Strassen's Theorem. Relationships between metrics.

Metric for convergence in probability. Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space, (S, d) - a metric space and $X, Y : \Omega \to S$ - random variables with values in S. The quantity

$$\alpha(X, Y) = \inf\{\varepsilon \geq 0 : \mathbb{P}(d(X, Y) > \varepsilon) \leq \varepsilon\}$$

is called the *Ky Fan metric* on the set $\mathcal{L}^0(\Omega, S)$ of classes of equivalences of such random variables, where two r.v.s are equivalent if they are equal a.s. If we take a sequence

$$\varepsilon_k \downarrow \alpha = \alpha(X, Y)$$

then $\mathbb{P}(d(X, Y) > \varepsilon_k) \leq \varepsilon_k$ and since

$$\mathrm{I}(d(X, Y) > \varepsilon_k) \uparrow \mathrm{I}(d(X, Y) > \alpha),$$

by monotone convergence theorem, $\mathbb{P}(d(X, Y) > \alpha) \leq \alpha$. Thus, the infimum in the definition of $\alpha(X, Y)$ is attained.

Lemma 37 *α is a metric on $\mathcal{L}^0(\Omega, S)$ which metrizes convergence in probability.*

Proof. First of all, clearly, $\alpha(X, Y) = 0$ iff $X = Y$ almost surely. To prove the triangle inequality,

$$\begin{aligned} \mathbb{P}(d(X, Z) > \alpha(X, Y) + \alpha(Y, Z)) &\leq \mathbb{P}(d(X, Y) > \alpha(X, Y)) + \mathbb{P}(d(Y, Z) > \alpha(Y, Z)) \\ &\leq \alpha(Y, Z) + \alpha(Y, Z) \end{aligned}$$

so that $\alpha(X, Z) \leq \alpha(X, Y) + \alpha(Y, Z)$. This proves that α is a metric. Next, if $\alpha_n = \alpha(X_n, X) \to 0$ then for any $\varepsilon > 0$ and large enough n such that $\alpha_n < \varepsilon$,

$$\mathbb{P}(d(X_n, X) > \varepsilon) \leq \mathbb{P}(d(X_n, X) > \alpha_n) \leq \alpha_n \to 0.$$

Conversely, if $X_n \to X$ in probability then for any $m \geq 1$ and large enough $n \geq n(m)$,

$$\mathbb{P}\Big(d(X_n, X) > \frac{1}{m}\Big) \leq \frac{1}{m}$$

which means that $\alpha_n \leq 1/m$ so that $\alpha_n \to 0$.

□

Lemma 38 *For $X, Y \in \mathcal{L}^0(\Omega, S)$, the Levy-Prohorov metric ρ satisfies*

$$\rho(\mathcal{L}(X), \mathcal{L}(Y)) \leq \alpha(X, Y).$$

Proof. Take $\varepsilon > \alpha(X,Y)$ so that $\mathbb{P}(d(X,Y) \geq \varepsilon) \leq \varepsilon$. For any set $A \subseteq S$,

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, d(X,Y) < \varepsilon) + \mathbb{P}(X \in A, d(X,Y) \geq \varepsilon) \leq \mathbb{P}(Y \in A^{\varepsilon}) + \varepsilon$$

which means that $\rho(\mathcal{L}(X), \mathcal{L}(Y)) \leq \varepsilon$. Letting $\varepsilon \downarrow \alpha(X,Y)$ proves the result.

□

We will now prove that, in some sense, the opposite is also true. Let (S,d) be a metric space and $\mathbb{P}, \mathbb{Q}$ be probability laws on S. Suppose that these laws are close in the Levy-Prohorov metric ρ. Can we construct random variables s_1 and s_2, with laws $\mathbb{P}$ and $\mathbb{Q}$, that are define on the same probability space and are close to each other in the Ky Fan metric α? We will construct a distribution on the product space $S \times S$ such that the coordinates s_1 and s_2 have marginal distributions $\mathbb{P}$ and $\mathbb{Q}$ and the distribution is concentrated in the neighborhood of the diagonal $s_1 = s_2$, where s_1 and s_2 are close in metric d, and the size of the neighborhood is controlled by $\rho(\mathbb{P}, \mathbb{Q})$.

Consider two sets X and Y. Given a subset $K \subseteq X \times Y$ and $A \subseteq X$ we define a K-image of A by

$$A^K = \{y \in Y : \exists x \in A, (x,y) \in K\}.$$

A *K-matching* f of X into Y is a one-to-one function $f : X \to Y$ such that $(x, f(x)) \in K$. We will need the following well known matching theorem.

Theorem 45 *If X, Y are finite and for all $A \subseteq X$,*

$$\text{card}(A^K) \geq \text{card}(A) \tag{19.0.1}$$

then there exists a K-matching f of X into Y.

Proof. We will prove the result by induction on $m = \text{card}(X)$. The case of $m = 1$ is obvious. For each $x \in X$ there exists $y \in Y$ such that $(x,y) \in K$. If there is a matching f of $X \setminus \{x\}$ into $Y \setminus \{y\}$ then defining $f(x) = y$ extends f to X. If not, then since $\text{card}(X \setminus \{x\}) < m$, by induction assumption, condition (19.0.1) is violated, i.e. there exists a set $A \subseteq X \setminus \{x\}$ such that $\text{card}(A^K \setminus \{y\}) < \text{card}(A)$. But because we also know that $\text{card}(A^K) \geq \text{card}(A)$ this implies that $\text{card}(A^K) = \text{card}(A)$. Since $\text{card}(A) < m$, by induction there exists a matching of A onto A^K. If there is a matching of $X \setminus A$ into $Y \setminus A^K$ we can combine it with a matching of A and A^K. If not, again by induction assumption, there exists $D \in X \setminus A$ such that $\text{card}(D^K \setminus A^K) < \text{card}(D)$. But then

$$\text{card}\Big((A \cup D)^K\Big) = \text{card}(D^K \setminus A^K) + \text{card}(A^K) < \text{card}(D) + \text{card}(A) = \text{card}(D \cup A),$$

which contradicts the assumption (19.0.1).

□

Theorem 46 *(Strassen) Suppose that (S,d) is a separable metric space and $\alpha, \beta > 0$. Suppose that laws $\mathbb{P}$ and $\mathbb{Q}$ are such that for all measurable sets $F \subseteq S$,*

$$\mathbb{P}(F) \leq \mathbb{Q}(F^{\alpha}) + \beta \tag{19.0.2}$$

Then for any $\varepsilon > 0$ there exist two non-negative measures η, γ on $S \times S$ such that

1. *$\mu = \eta + \gamma$ is a law on $S \times S$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.*
2. *$\eta(d(x,y) > \alpha + \varepsilon) = 0$.*
3. *$\gamma(S \times S) \leq \beta + \varepsilon$.*
4. *μ is a finite sum of product measures.*

Remark. Condition (19.0.2) is a relaxation of the definition of the Levy-Prohorov metric, one can take any $\alpha, \beta > \rho(\mathbb{P}, \mathbb{Q})$. Conditions 1 - 3 mean that we can construct a measure μ on $S \times S$ such that coordinates x, y have marginal distributions $\mathbb{P}, \mathbb{Q}$, concentrated within distance $\alpha + \varepsilon$ of each other (condition 2) except for the set of measure at most $\beta + \varepsilon$ (condition 3).

Proof. The proof will proceed in several steps.

Case A. We will start with the simplest case which is, however, at the core of everything else. Given small $\varepsilon > 0$, take $n \geq 1$ such that $n\varepsilon > 1$. Suppose that laws $\mathbb{P}, \mathbb{Q}$ are uniform on finite subsets $M, N \subseteq S$ of equal cardinality,

$$\text{card}(M) = \text{card}(N) = n, \;\; \mathbb{P}(x) = \mathbb{Q}(y) = \frac{1}{n} < \varepsilon, \;\; x \in M, y \in N.$$

Using condition (19.0.2), we would like to match as many points from M and N as possible, but only points that are within distance α from each other. To use the matching theorem, we will introduce some auxiliary sets U and V that are not too big, with size controlled by parameter β, and the union of these sets with M and N satisfies a certain matching condition.

Take integer k such that $\beta n \leq k < (\beta + \varepsilon)n$. Let us take sets U and V such that $k = \text{card}(U) = \text{card}(V)$ and U, V are disjoint from M, N. Define

$$X = M \cup U, \;\; Y = N \cup V.$$

Let us define a subset $K \subseteq X \times Y$ such that $(x, y) \in K$ if and only if one of the following holds:

1. $x \in U$,
2. $y \in V$,
3. $d(x, y) \leq \alpha$ if $x \in M, y \in N$.

This means that small auxiliary sets can be matched with any points but only close points, $d(x, y) \leq \alpha$, can be matched in the main sets M and N. Consider a set $A \subseteq X$ with cardinality $\text{card}(A) = r$. If $A \not\subseteq M$ then by 1, $A^K = Y$ and $\text{card}(A^K) \geq r$. Suppose now that $A \subseteq M$ and we would like to show that again $\text{card}(A^K) \geq r$. By (19.0.2),

$$\frac{r}{n} = \mathbb{P}(A) \leq \mathbb{Q}(A^\alpha) + \beta = \frac{1}{n}\text{card}(A^\alpha \cap N) + \beta \leq \frac{1}{n}\text{card}(A^K \cap N) + \beta$$

since by 3, $A^\alpha \subseteq A^K$. Therefore,

$$r = \text{card}(A) \leq n\beta + \text{card}(A^K \cap N) \leq k + \text{card}(A^K \cap N) = \text{card}(A^K),$$

since $k = \text{card}(V)$ and $A^K = V \cup (A^K \cap N)$. By matching theorem, there exists a K-matching f of X and Y. Let

$$T = \{x \in M : f(x) \in N\},$$

i.e. close points, $d(x, y) \leq \alpha$, from M that are matched with points in N. Clearly, $\text{card}(T) \geq n - k$ and for $x \in T$, by 3, $d(x, f(x)) \leq \alpha$. For $x \in M \setminus T$, redefine $f(x)$ to match x with arbitrary points in N that are not matched with points in T. This defines a matching of M onto N. We define measures η and γ by

$$\eta = \frac{1}{n}\sum_{x \in T} \delta(x, f(x)), \;\; \gamma = \frac{1}{n}\sum_{x \in M \setminus T} \delta(x, f(x)),$$

and let $\mu = \eta + \gamma$. First of all, obviously, μ has marginals $\mathbb{P}$ and $\mathbb{Q}$ because each point in M or N appears in the sum $\eta + \gamma$ only once with weight $1/n$. Also,

$$\eta(d(x, f(x)) > \alpha) = 0, \;\; \gamma(S \times S) \leq \frac{\text{card}(M \setminus T)}{n} \leq \frac{k}{n} < \beta + \varepsilon. \tag{19.0.3}$$

Finally, both η and γ are finite sums of point masses which are product measures of point masses.

Case B. Suppose now that $\mathbb{P}$ and $\mathbb{Q}$ are concentrated on finitely many points with rational probabilities. Then we can artificially split all points into "smaller" points of equal probabilities as follows. Let n be such that $n\varepsilon > 1$ and

$$n\mathbb{P}(x), n\mathbb{Q}(x) \in J = \{1, 2, \ldots, n\}.$$

Define a discrete metric on J by $f(i,j) = \varepsilon\, \mathrm{I}(i \neq j)$ and define a metric on $S \times J$ by

$$e\big((x,i),(y,j)\big) = d(x,y) + f(i,j).$$

Define a measure $\mathbb{P}'$ on $S \times J$ as follows. If $\mathbb{P}(x) = \frac{j}{n}$ then

$$\mathbb{P}'\big((x,i)\big) = \frac{1}{n} \quad \text{for} \quad i = 1, \ldots, j.$$

Define $\mathbb{Q}'$ similarly. Let us check that laws $\mathbb{P}', \mathbb{Q}'$ satisfy the assumptions of Case A. Given a set $F \subseteq S \times J$, define

$$F_1 = \{x \in S : (x,j) \in F \text{ for some } j\}.$$

Using (19.0.2),

$$\mathbb{P}'(F) \leq \mathbb{P}(F_1) \leq \mathbb{Q}(F_1^{\alpha}) + \beta \leq \mathbb{Q}'(F^{\alpha+\varepsilon}) + \beta,$$

because $f(i,j) \leq \varepsilon$. By Case A in (19.0.3), we can construct $\mu' = \eta' + \gamma'$ with marginals $\mathbb{P}'$ and $\mathbb{Q}'$ such that

$$\eta'\big(e((x,i),(y,j)) > \alpha + \varepsilon\big) = 0, \;\; \gamma'\big((S \times J) \times (S \times J)\big) < \beta + \varepsilon.$$

Let μ, η, γ be the projections of μ', η', γ' back onto $S \times S$ by the map $((x,i),(y,j)) \to (x,y)$. Then, clearly, $\mu = \eta + \gamma$, μ has marginals $\mathbb{P}$ and $\mathbb{Q}$ and $\gamma(S \times S) < \beta + \varepsilon$. Finally, since

$$e\big((x,i),(y,j)\big) = d(x,y) + f(i,j) \geq d(x,y),$$

we get

$$\eta\big(d(x,y) > \alpha + \varepsilon\big) \leq \eta'\big(e((x,i),(y,j)) > \alpha + \varepsilon\big) = 0.$$

Case C. (General case) Let $\mathbb{P}, \mathbb{Q}$ be the laws on a separable metric space (S,d). Let A be a maximal set such that for all $x, y \in A, d(x,y) \geq \varepsilon$. The set A is countable, $A = \{x_i\}_{i \geq 1}$, because S is separable, and since A is maximal, for all $x \in S$ there exists $y \in A$ such that $d(x,y) < \varepsilon$. Such set A is usually called an ε-packing. Let us create a partition of S using ε-balls around $\{x_i\}$:

$$B_1 = \{x \in S : d(x, x_1) < \varepsilon\}, \;\; B_2 = \{d(x, x_2) < \varepsilon\} \setminus B_1$$

and, iteratively for $k \geq 2$,

$$B_k = \{d(x, x_k) < \varepsilon\} \setminus (B_1 \cup \cdots \cup B_{k-1}).$$

$\{B_k\}_{k \geq 1}$ is a partition of S. Let us discretize measures $\mathbb{P}$ and $\mathbb{Q}$ by projecting them onto $\{x_i\}_{i \geq 1}$:

$$\mathbb{P}'(x_k) = \mathbb{P}(B_k), \;\; \mathbb{Q}'(x_k) = \mathbb{Q}(B_k).$$

Consider any set $F \subseteq S$. For any point $x \in F$, if $x \in B_k$ then $d(x, x_k) < \varepsilon$, i.e. $x_k \in F^{\varepsilon}$ and, therefore,

$$\mathbb{P}(F) \leq \mathbb{P}'(F^{\varepsilon}).$$

Also, if $x_k \in F$ then $B_k \subseteq F^{\varepsilon}$ and, therefore,

$$\mathbb{P}'(F) \leq \mathbb{P}(F^{\varepsilon}).$$

To apply Case B, we need to approximate $\mathbb{P}'$ by a measure on a finite number of points with rationals probabilities. For large enough $n \geq 1$, let

$$\mathbb{P}''(x_k) = \frac{\lfloor n\mathbb{P}'(x_k) \rfloor}{n}.$$

Clearly, as $n \to \infty$, $\mathbb{P}''(x_k) \uparrow \mathbb{P}'(x_k)$. Since only a finite number of points carry non-zero weights $\mathbb{P}''(x_k) > 0$, let x_0 be one of the other points in the sequence $\{x_k\}$. Let us assign to it a probability

$$\mathbb{P}''(x_0) = 1 - \sum_{k \geq 1} \mathbb{P}''(x_k).$$

If we take n large enough so that $\mathbb{P}''(x_0) < \varepsilon/2$ then

$$\sum_{k \geq 0} |\mathbb{P}''(x_k) - \mathbb{P}'(x_k)| \leq \varepsilon.$$

All the relations above also hold true for $\mathbb{Q}, \mathbb{Q}'$ and $\mathbb{Q}''$ that are defined similarly. We can write for $F \subseteq S$

$$\mathbb{P}''(F) \leq \mathbb{P}'(F) + \varepsilon \leq \mathbb{P}(F^{\varepsilon}) + \varepsilon \leq \mathbb{Q}(F^{\varepsilon+\alpha}) + \beta + \varepsilon \leq \mathbb{Q}'(F^{\alpha+2\varepsilon}) + \beta + \varepsilon \leq \mathbb{Q}''(F^{\alpha+2\varepsilon}) + \beta + 2\varepsilon.$$

By Case B, there exists a decomposition $\mu'' = \eta'' + \gamma''$ on $S \times S$ with marginals $\mathbb{P}''$ and $\mathbb{Q}''$ such that

$$\eta''\big(d(x,y) > \alpha + 3\varepsilon\big) = 0, \;\; \gamma''(S \times S) \leq \beta + 3\varepsilon.$$

Let us also assume that the points (x_0, x_i) and (x_i, x_0) for $i \geq 0$ are included in the support of γ''. Since the total weight of these points is at most ε, the total weight of γ'' does no increase much:

$$\gamma''(S \times S) \leq \beta + 5\varepsilon.$$

It remains to redistribute these measures from sequence $\{x_i\}_{i \geq 0}$ to S in a way that recovers marginal distributions $\mathbb{P}$ and $\mathbb{Q}$ and so that not much accuracy is lost. Define a sequence of measures on S by

$$\mathbb{P}_i(C) = \frac{\mathbb{P}(CB_i)}{\mathbb{P}(B_i)} \;\; \text{if } \mathbb{P}(B_i) > 0 \text{ and } \mathbb{P}_i(C) = 0 \text{ otherwise}$$

and define $\mathbb{Q}_i$ similarly. The measures $\mathbb{P}_i$ and $\mathbb{Q}_i$ are concentrated on B_i. Define

$$\eta = \sum_{i,j \geq 1} \eta''(x_i, x_j)(\mathbb{P}_i \times \mathbb{Q}_j)$$

The marginals of η satisfy

$$\begin{aligned} u(C) = \eta(C \times S) \;\; &\leq \;\; \sum_{i,j \geq 1} \eta''(x_i, x_j)\mathbb{P}_i(C) = \sum_{i \geq 1} \eta''(x_i, S)\mathbb{P}_i(C) \\ &\leq \;\; \sum_{i \geq 1} \mathbb{P}''(x_i)\mathbb{P}_i(C) \leq \sum_{i \geq 1} \mathbb{P}'(x_i)\mathbb{P}_i(C) = \sum_{i \geq 1} \mathbb{P}(B_i)\mathbb{P}_i(C) = \mathbb{P}(C) \end{aligned}$$

and, similarly,

$$v(C) = \eta(S \times C) \leq \mathbb{Q}(C).$$

Since $\eta''(x_i, x_j) = 0$ unless $d(x_i, x_j) \leq \alpha + 3\varepsilon$, the measure

$$\eta = \sum_{i,j \geq 1} \eta''(x_i, x_j)(\mathbb{P}_i \times \mathbb{Q}_j)$$

is concentrated on the set $\{d(x,y) \leq \alpha + 5\varepsilon\}$ because for $x \in B_i, y \in B_j$,

$$d(x,y) \leq d(x, x_i) + d(x_i, x_j) + d(x_j, y) \leq \varepsilon + \alpha + 3\varepsilon + \varepsilon = \alpha + 5\varepsilon.$$

If $u(S) = v(S) = 1$ then $\eta(S \times S) = 1$ and η has marginals $\mathbb{P}$ and $\mathbb{Q}$ so we can take $\gamma = 0$. Otherwise, take $t = 1 - u(S)$ and define

$$\gamma = \frac{1}{t}(\mathbb{P} - u) \times (\mathbb{Q} - v).$$

It is easy to check that $\mu = \eta + \gamma$ has marginals $\mathbb{P}$ and $\mathbb{Q}$. Also,

$$\gamma(S \times S) = t = 1 - \eta(S \times S) = 1 - \eta''(S \times S) = \gamma''(S \times S) \leq \beta + 5\varepsilon.$$

□

Relationships between metrics. The following relationship between Ky Fan and Levy-Prohorov metrics is an immediate consequence of Strassen's theorem. We already saw that $\rho(\mathcal{L}(X), \mathcal{L}(Y)) \leq \alpha(X, Y)$.

Theorem 47 *If (S, d) is a separable metric space and $\mathbb{P}, \mathbb{Q}$ are laws on S then for any $\varepsilon > 0$ there exist random variables X and Y with distributions $\mathcal{L}(X) = \mathbb{P}$ and $\mathcal{L}(Y) = \mathbb{Q}$ such that*

$$\alpha(X, Y) \leq \rho(\mathbb{P}, \mathbb{Q}) + \varepsilon.$$

If $\mathbb{P}$ and $\mathbb{Q}$ are tight, one can take $\varepsilon = 0$.

Proof. Let us take $\alpha = \beta = \rho(\mathbb{P}, \mathbb{Q})$. Then, by definition of the Levy-Prohorov metric, for any $\varepsilon > 0$ and for any set A,

$$\mathbb{P}(A) \leq \mathbb{Q}(A^{\rho+\varepsilon}) + \rho + \varepsilon.$$

By Strassen's theorem, there exists a measure μ on $S \times S$ with marginals $\mathbb{P}, \mathbb{Q}$ such that

$$\mu\big(d(x, y) > \rho + 2\varepsilon\big) \leq \rho + 2\varepsilon. \tag{19.0.4}$$

Therefore, if X and Y are the coordinates of $S \times S$, i.e.

$$X, Y : S \times S \to S, \;\; X(x, y) = x, \;\; Y(x, y) = y,$$

then by definition of the Ky Fan metric, $\alpha(X, Y) \leq \rho + 2\varepsilon$. If $\mathbb{P}$ and $\mathbb{Q}$ are tight then there exists a compact K such that $\mathbb{P}(K), \mathbb{Q}(K) \geq 1 - \delta$. For $\varepsilon = 1/n$ find μ_n as in (19.0.4). Since μ_n has marginals $\mathbb{P}$ and $\mathbb{Q}$, $\mu_n(K \times K) \geq 1 - 2\delta$, which means that $(\mu_n)_{n\geq 1}$ are uniformly tight. By selection theorem, there exists a convergent subsequence $\mu_{n(k)} \to \mu$. Obviously, μ has marginals $\mathbb{P}$ and $\mathbb{Q}$. Since by construction,

$$\mu_n\Big(d(x, y) > \rho + \frac{2}{n}\Big) \leq \rho + \frac{2}{n}$$

and $\{d(x, y) > \rho + 2/n\}$ is an open set on $S \times S$, by portmanteau theorem,

$$\mu\Big(d(x, y) > \rho + \frac{2}{n}\Big) \leq \liminf_{k\to\infty} \mu_{n(k)}\Big(d(x, y) > \rho + \frac{2}{n(k)}\Big) \leq \rho.$$

Letting $n \to \infty$ we get $\mu(d(x, y) > \rho) \leq \rho$ and, therefore, $\alpha(X, Y) \leq \rho$.

□

This also implies the relationship between the Bounded Lipschitz metric β and Levy-Prohorov metric ρ.

Lemma 39 *If (S, d) is a separable metric space then*

$$\frac{1}{2}\beta(\mathbb{P}, \mathbb{Q}) \leq \rho(\mathbb{P}, \mathbb{Q}) \leq 2\sqrt{\beta(\mathbb{P}, \mathbb{Q})}.$$

Proof. We already proved the second inequality. To prove the first one, given $\varepsilon > 0$ take random variables X and Y such that $\alpha(X, Y) \leq \rho + \varepsilon$. Consider a bounded Lipschitz function f, $||f||_{\mathrm{BL}} < \infty$. Then

$$\begin{aligned}
\Big|\int f d\mathbb{P} - \int f d\mathbb{Q}\Big| &= |\mathbb{E}f(X) - \mathbb{E}f(Y)| \leq \mathbb{E}|f(X) - f(Y)| \\
&\leq \|f\|_{\mathrm{L}}(\rho + \varepsilon) + 2\|f\|_\infty \mathbb{P}\Big(d(X, Y) > \rho + \varepsilon\Big) \\
&\leq \|f\|_{\mathrm{L}}(\rho + \varepsilon) + 2\|f\|_\infty(\rho + \varepsilon) \leq 2\|f\|_{\mathrm{BL}}(\rho + \varepsilon).
\end{aligned}$$

Thus, $\beta(\mathbb{P}, \mathbb{Q}) \leq 2(\rho(\mathbb{P}, \mathbb{Q}) + \varepsilon)$ and letting $\varepsilon \to 0$ finishes the proof.

□

Section 20

Kantorovich-Rubinstein Theorem.

Let (S, d) be a separable metric space. Denote by $\mathcal{P}_1(S)$ the set of all laws on S such that for some $z \in S$ (equivalently, for all $z \in S$),

$$\int_S d(x, z)\mathbb{P}(x) < \infty.$$

Let us denote by

$$M(\mathbb{P}, \mathbb{Q}) = \Big\{\mu \,:\, \mu \text{ is a law on } S \times S \text{ with marginals } \mathbb{P} \text{ and } \mathbb{Q}\Big\}.$$

Definition. For $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_1(S)$, the quantity

$$W(\mathbb{P}, \mathbb{Q}) = \inf\Big\{\int d(x, y)d\mu(x, y) : \mu \in M(\mathbb{P}, \mathbb{Q})\Big\}$$

is called the *Wasserstein* distance between $\mathbb{P}$ and $\mathbb{Q}$.

A measure $\mu \in M(\mathbb{P}, \mathbb{Q})$ represents a *transportation* between measures $\mathbb{P}$ and $\mathbb{Q}$. We can think of the conditional distribution $\mu(y|x)$ as a way to redistribute the mass in the neighborhood of a point x so that the distribution $\mathbb{P}$ will be redistributed to the distribution $\mathbb{Q}$. If the distance $d(x, y)$ represents the cost of moving x to y then the Wasserstein distance gives the optimal total cost of transporting $\mathbb{P}$ to $\mathbb{Q}$.

Given any two laws $\mathbb{P}$ and $\mathbb{Q}$ on S, let us define

$$\gamma(\mathbb{P}, \mathbb{Q}) = \sup\Big\{\Big|\int f d\mathbb{P} - \int f d\mathbb{Q}\Big| : ||f||_{\mathrm{L}} \le 1\Big\}$$

and

$$m_d(\mathbb{P}, \mathbb{Q}) = \sup\Big\{\int f d\mathbb{P} + \int g d\mathbb{Q} \,:\, f, g \in C(S),\ f(x) + g(y) < d(x, y)\Big\}.$$

Lemma 40 *We have* $\gamma(\mathbb{P}, \mathbb{Q}) = m_d(\mathbb{P}, \mathbb{Q})$.

Proof. Given a function f such that $||f||_{\mathrm{L}} \le 1$ let us take a small $\varepsilon > 0$ and $g(y) = -f(y) - \varepsilon$. Then

$$f(x) + g(y) = f(x) - f(y) - \varepsilon \le d(x, y) - \varepsilon < d(x, y)$$

and

$$\int f d\mathbb{P} + \int g d\mathbb{Q} = \int f d\mathbb{P} - \int f d\mathbb{Q} - \varepsilon.$$

Combining with the choice of $-f(x)$ and $g(y) = f(y) - \varepsilon$ we get

$$\Big|\int f d\mathbb{P} - \int f d\mathbb{Q}\Big| \le \sup\Big\{\int f d\mathbb{P} + \int g d\mathbb{Q} : f(x) + g(y) < d(x, y)\Big\} + \varepsilon$$

which, of course, proves that

$$\gamma(\mathbb{P},\mathbb{Q}) \leq \sup\Big\{\int f d\mathbb{P} + \int g d\mathbb{Q} : f(x)+g(y) < d(x,y)\Big\}.$$

Let us now consider functions f, g such that $f(x)+g(y) < d(x,y)$. Define

$$e(x) = \inf_y (d(x,y) - g(y)) = -\sup_y (g(y) - d(x,y))$$

Clearly,

$$f(x) \leq e(x) \leq d(x,x) - g(x) = -g(x)$$

and, therefore,

$$\int f d\mathbb{P} + \int g d\mathbb{Q} \leq \int e d\mathbb{P} - \int e d\mathbb{Q}.$$

Function e satisfies

$$\begin{aligned} e(x) - e(x') &= \sup_y \big(g(y) - d(x',y)\big) - \sup_y \big(g(y) - d(x,y)\big) \\ &\leq \sup_y \big(d(x,y) - d(x',y)\big) \leq d(x,x') \end{aligned}$$

which means that $||e||_{\mathrm{L}} = 1$. This finishes the proof.

□

We will need the following version of the Hahn-Banach theorem.

Theorem 48 *(Hahn-Banach) Let V be a normed vector space, E - a linear subspace of V and U - an open convex set in V such that $U \cap E \neq \emptyset$. If $r : E \to \mathbb{R}$ is a linear non-zero functional on E then there exists a linear functional $\rho : V \to \mathbb{R}$ such that $\rho|_E = r$ and $\sup_U \rho(x) = \sup_{U\cap E} r(x)$.*

Proof. Let $t = \sup\{r(x) : x \in U \cap E\}$ and let $B = \{x \in E : r(x) > t\}$. Since B is convex and $U \cap B = \emptyset$, the Hahn-Banach separation theorem implies that there exists a linear functional $q : V \to \mathbb{R}$ such that $\sup_U q(x) \leq \inf_B q(x)$. For any $x_0 \in U \cap E$ let $F = \{x \in E : q(x) = q(x_0)\}$. Since $q(x_0) < \inf_B q(x)$, $F \cap B = \emptyset$. This means that the hyperplanes $\{x \in E : q(x) = q(x_0)\}$ and $\{x \in E : r(x) = t\}$ in the subspace E are parallel and this implies that $q(x) = \alpha r(x)$ on E for some $\alpha \neq 0$. Let $\rho = q/\alpha$. Then $r = \rho|_E$ and

$$\sup_U \rho(x) = \frac{1}{\alpha} \sup_U q(x) \leq \frac{1}{\alpha} \inf_B q(x) = \inf_B r(x) = t = \sup_{U\cap E} r(x).$$

Since $r = \rho|_E$, this finishes the proof.

□

Theorem 49 *If S is a compact metric space then $W(\mathbb{P},\mathbb{Q}) = m_d(\mathbb{P},\mathbb{Q})$ for $\mathbb{P},\mathbb{Q} \in \mathcal{P}_1(S)$.*

Proof. Consider a vector space $V = C(S \times S)$ equipped with $\|\cdot\|_\infty$ norm and let

$$U = \big\{f \in V : f(x,y) < d(x,y)\big\}.$$

Obviously, U is convex and open because $S \times S$ is compact and any continuous function on a compact achieves its maximum. Consider a linear subspace E of V defined by

$$E = \big\{\phi \in V \,:\, \phi(x,y) = f(x) + g(y)\big\}$$

so that

$$U \cap E = \big\{f(x) + g(y) < d(x,y)\big\}.$$

Define a linear functional r on E by

$$r(\phi) = \int f d\mathbb{P} + \int g d\mathbb{Q} \quad \text{if} \quad \phi = f(x) + g(y).$$

By the above Hahn-Banach theorem, r can be extended to $\rho : V \to \mathbb{R}$ such that $\rho|_E = r$ and

$$\sup_U \rho(\phi) = \sup_{U \cap E} r(\phi) = m_d(\mathbb{P}, \mathbb{Q}).$$

Let us look at the properties of this functional. First of all, if $a(x, y) \geq 0$ then $\rho(a) \geq 0$. Indeed, for any $c \geq 0$

$$U \ni d(x, y) - c \cdot a(x, y) - \varepsilon < d(x, y)$$

and, therefore, for all $c \geq 0$

$$\rho(d - ca - \varepsilon) = \rho(d) - c\rho(a) - \rho(\varepsilon) \leq \sup_U \rho < \infty.$$

This can hold only if $\rho(a) \geq 0$. This implies that if $\phi_1 \leq \phi_2$ then $\rho(\phi_1) \leq \rho(\phi_2)$. For any function ϕ, both $-\phi, \phi \leq \|\phi\|_\infty \cdot 1$ and, by monotonicity of ρ,

$$|\rho(\phi)| \leq ||\phi||_\infty \rho(1) = ||\phi||_\infty.$$

Since $S \times S$ is compact and ρ is a continuous functional on $(C(S \times S), \|\cdot\|_\infty)$, by the Reisz representation theorem there exists a unique measure μ on the Borel σ-algebra on $S \times S$ such that

$$\rho(f) = \int f(x, y) d\mu(x, y).$$

Since $\rho|_E = r$,

$$\int (f(x) + g(y)) d\mu(x, y) = \int f d\mathbb{P} + \int g d\mathbb{Q}$$

which implies that $\mu \in M(\mathbb{P}, \mathbb{Q})$. We have

$$m_d(\mathbb{P}, \mathbb{Q}) = \sup_U \rho(\phi) = \sup \left\{ \int f(x, y) d\mu(x, y) : f(x, y) < d(x, y) \right\} = \int d(x, y) d\mu(x, y) \geq W(\mathbb{P}, \mathbb{Q}).$$

The opposite inequality is easy because for any f, g such that $f(x) + g(y) < d(x, y)$ and any $\nu \in M(\mathbb{P}, \mathbb{Q})$,

$$\int f d\mathbb{P} + \int g d\mathbb{Q} = \int (f(x) + g(y)) d\nu(x, y) \leq \int d(x, y) d\nu(x, y). \tag{20.0.1}$$

This finishes the proof and, moreover, it shows that the infimum in the definition of W is achieved on μ. □

Remark. Notice that in the proof of this theorem we never used the fact that d is a metric. Theorem holds for any $d \in C(S \times S)$ under the corresponding integrability assumptions. For example, one can consider loss functions of the type $d(x, y)^p$ for $p > 1$, which are not necessarily metrics. However, in Lemma 40, the fact that d is a metric was essential.

Our next goal will be to show that $W = \gamma$ on separable and not necessarily compact metric spaces. We start with the following.

Lemma 41 *If (S, d) is a separable metric space then W and γ are metrics on $\mathcal{P}_1(S)$.*

Proof. Since for a bounded Lipschitz metric β we have $\beta(\mathbb{P}, \mathbb{Q}) \leq \gamma(\mathbb{P}, \mathbb{Q})$, γ is also a metric because if $\gamma(\mathbb{P}, \mathbb{Q}) = 0$ then $\beta(\mathbb{P}, \mathbb{Q}) = 0$ and, therefore, $\mathbb{P} = \mathbb{Q}$. As in (20.0.1), it should be obvious that $\gamma(\mathbb{P}, \mathbb{Q}) = m_d(\mathbb{P}, \mathbb{Q}) \leq W(\mathbb{P}, \mathbb{Q})$ and if $W(\mathbb{P}, \mathbb{Q}) = 0$ then $\gamma(\mathbb{P}, \mathbb{Q}) = 0$ and $\mathbb{P} = \mathbb{Q}$. Symmetry of W is obvious. It remains to show that $W(\mathbb{P}, \mathbb{Q})$ satisfies the triangle inequality. The idea will be rather simple, but to have well-defined

conditional distributions we will need to approximate distributions on $S \times S$ with given marginals by a more regular disributions with the same marginals. Let us first explain the main idea. Consider three laws $\mathbb{P}, \mathbb{Q}, \mathbb{T}$ on S and let $\mu \in M(\mathbb{P}, \mathbb{Q})$ and $\nu \in M(\mathbb{Q}, \mathbb{T})$ be such that

$$\int d(x,y)d\mu(x,y) \le W(\mathbb{P},\mathbb{Q}) + \varepsilon \ \text{ and } \ \int d(y,z)d\nu(y,z) \le W(\mathbb{Q},\mathbb{T}) + \varepsilon.$$

Let us generate a distribution γ on $S \times S \times S$ with marginals $\mathbb{P}, \mathbb{Q}$ and $\mathbb{T}$ and marginals on pairs of coordinates (x,y) and (y,z) given by μ and ν by "gluing" μ and ν in the following way. Let us generate y from distribution $\mathbb{Q}$ and, given y, generate x and z according to conditional distributions $\mu(x|y)$ and $\nu(z|y)$ independently of each other, i.e.

$$\gamma(x,z|y) = \mu(x|y) \times \nu(z|y).$$

Obviously, by construction, (x,y) has distribution μ and (y,z) has distribution ν. Therefore, the marginals of x and z are $\mathbb{P}$ and $\mathbb{T}$ which means that the pair (x,z) has distribution $\eta \in M(\mathbb{P},\mathbb{T})$. Finally,

$$\begin{aligned} W(\mathbb{P},\mathbb{T}) &\le \int d(x,z)d\eta(x,z) = \int d(x,z)d\gamma(x,y,z) \le \int d(x,y)d\gamma + \int d(y,z)d\gamma \\ &= \int d(x,y)d\mu + \int d(y,z)d\nu \le W(\mathbb{P},\mathbb{Q}) + W(\mathbb{Q},\mathbb{T}) + 2\varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0$ proves the triangle inequality for W. It remains to explain how the conditional distributions can be well defined. Let us modify μ by 'discretizing' it without losing much in the transportation cost integral. Given $\varepsilon > 0$, consider a partition $(S_n)_{n\ge 1}$ of S such that diameter$(S_n) < \varepsilon$ for all n. This can be done as in the proof of Strassen's theorem, Case C. On each box $S_n \times S_m$ let

$$\mu^1_{nm}(C) = \frac{\mu((C \cap S_n) \times S_m)}{\mu(S_n \times S_m)}, \ \ \mu^2_{nm}(C) = \frac{\mu(S_n \times (C \cap S_m))}{\mu(S_n \times S_m)}$$

be the marginal distributions of the conditional distribution of μ on $S_n \times S_m$. Define

$$\mu' = \sum_{n,m} \mu(S_n \times S_m)\ \mu^1_{nm} \times \mu^2_{nm}.$$

In this construction, locally on each small box $S_n \times S_m$, measure μ is replaced by the product measure with the same marginals. Let us compute the marginals of μ'. Given a set $C \subseteq S$,

$$\begin{aligned} \mu'(C \times S) &= \sum_{n,m} \mu(S_n \times S_m)\ \mu^1_{nm}(C) \times \mu^2_{nm}(S) \\ &= \sum_{n,m} \mu((C \cap S_n) \times S_m) = \sum_n \mu((C \cap S_n) \times S) = \sum_n \mathbb{P}(C \cap S_n) = \mathbb{P}(C). \end{aligned}$$

Similarly, $\mu'(S \times C) = \mathbb{Q}(C)$, so μ' has the same marginals as μ, $\mu' \in M(\mathbb{P},\mathbb{Q})$. It should be obvious that transportation cost integral does not change much by replacing μ with μ'. One can visualize this by looking at what happens locally on each small box $S_n \times S_m$. Let (X_n, Y_m) be a random pair with distribution μ restricted to $S_n \times S_m$ so that

$$\mathbb{E}d(X_n, Y_m) = \frac{1}{\mu(S_n \times S_m)} \int_{S_n \times S_m} d(x,y)d\mu(x,y).$$

Let Y'_m be an independent copy of Y_m, also independent of X_n, i.e. the joint distribution of (X_n, Y'_m) is $\mu^1_{nm} \times \mu^2_{nm}$ and

$$\mathbb{E}d(X_n, Y'_m) = \int_{S_n \times S_m} d(x,y)d(\mu^1_{nm} \times \mu^2_{nm})(x,y).$$

Then

$$\int d(x,y)d\mu(x,y) = \sum_{n,m} \mu(S_n \times S_m)\mathbb{E}d(X_n, Y_m),$$

$$\int d(x,y)d\mu'(x,y) = \sum_{n,m} \mu(S_n \times S_m)\mathbb{E}d(X_n, Y'_m).$$

Finally, $d(Y_m, Y'_m) \leq \text{diam}(S_m) \leq \varepsilon$ and these two integrals differ by at most ε. Therefore,

$$\int d(x,y)d\mu'(x,y) \leq W(\mathbb{P},\mathbb{Q}) + 2\varepsilon.$$

Similarly, we can define

$$\nu' = \sum_{n,m} \nu(S_n \times S_m)\ \nu^1_{nm} \times \nu^2_{nm}$$

such that

$$\int d(x,y)d\nu'(x,y) \leq W(\mathbb{Q},\mathbb{T}) + 2\varepsilon.$$

We will now show that this special simple form of the distributions $\mu'(x,y), \nu'(y,z)$ ensures that the conditional distributions of x and z given y are well defined. Let $\mathbb{Q}_m$ be the restriction of $\mathbb{Q}$ to S_m,

$$\mathbb{Q}_m(C) = \mathbb{Q}(C \cap S_m) = \sum_n \mu(S_n \times S_m)\ \mu^2_{nm}(C).$$

Obviously, if $\mathbb{Q}_m(C) = 0$ then $\mu^2_{nm}(C) = 0$ for all n, which means that μ^2_{nm} are absolutely continuous with respect to $\mathbb{Q}_m$ and the Radon-Nikodym derivatives

$$f_{nm}(y) = \frac{d\mu^2_{nm}}{d\mathbb{Q}_m}(y) \ \text{ exist and } \ \sum_n \mu(S_n \times S_m) f_{nm}(y) = 1 \ \text{ a.s. for } y \in S_m.$$

Let us define a conditional distribution of x given y by

$$\mu'(A|y) = \sum_{n,m} \mu(S_n \times S_m) f_{nm}(y) \mu^1_{nm}(A).$$

Notice that for any $A \in \mathcal{B}$, $\mu'(A|y)$ is measurable in y and $\mu'(A|y)$ is a probability distribution on $\mathcal{B}$, $\mathbb{Q}$-a.s. over y because

$$\mu'(S|y) = \sum_{n,m} \mu(S_n \times S_m) f_{nm}(y) = 1 \ a.s.$$

Let us check that for Borel sets $A, B \in \mathcal{B}$,

$$\mu'(A \times B) = \int_B \mu'(A|y) d\mathbb{Q}(y).$$

Indeed, since $f_{nm}(y) = 0$ for $y \notin S_m$,

$$\begin{aligned}
\int_B \mu'(A|y)d\mathbb{Q}(y) &= \sum_{n,m} \mu(S_n \times S_m)\mu^1_{nm}(A) \int_B f_{nm}(y)d\mathbb{Q}(y) \\
&= \sum_{n,m} \mu(S_n \times S_m)\mu^1_{nm}(A) \int_B f_{nm}(y)d\mathbb{Q}_m(y) \\
&= \sum_{n,m} \mu(S_n \times S_m)\mu^1_{nm}(A)\mu^2_{nm}(B) = \mu'(A \times B).
\end{aligned}$$

Conditional distribution $\nu'(\cdot|y)$ can be defined similarly.

□

Next lemma shows that on a separable metric space any law with the "first moment", i.e. $\mathbb{P} \in \mathcal{P}_1(S)$, can be approximated in metrics W and γ by laws concentrated on finite sets.

Lemma 42 *If (S,d) is separable and $\mathbb{P}\in\mathcal{P}_1(S)$ then there exists a sequence of laws $\mathbb{P}_n$ such that $\mathbb{P}_n(F_n)=1$ for some finite sets F_n and $W(\mathbb{P}_n,\mathbb{P}),\gamma(\mathbb{P}_n,\mathbb{P})\to 0$.*

Proof. For each $n\geq 1$, let $(S_{nj})_{j\geq 1}$ be a partition of S such that $\mathrm{diam}(S_{nj})\leq 1/n$. Take a point $x_{nj}\in S_{nj}$ in each set S_{nj} and for $k\geq 1$ define a function

$$f_{nk}(x)=\begin{cases} x_{nj}, & \text{if } x\in S_{nj} \text{ for } j\leq k,\\ x_{n1}, & \text{if } x\in S_{nj} \text{ for } j>k.\end{cases}$$

We have,

$$\int d(x,f_{nk}(x))d\mathbb{P}(x)=\sum_{j\geq 1}\int_{S_{nj}} d(x,f_{nk}(x))d\mathbb{P}(x)\leq \frac{1}{n}\sum_{j\leq k}\mathbb{P}(S_{nj})+\int_{S\setminus(S_{n1}\cup\cdots\cup S_{nk})} d(x,x_{n1})d\mathbb{P}(x)\leq\frac{2}{n}$$

for k large enough because $\mathbb{P}\in\mathcal{P}_1(S)$, i.e. $\int d(x,x_{n1})d\mathbb{P}(x)<\infty$, and the set $S\setminus(S_{n1}\cup\cdots\cup S_{nk})\downarrow\emptyset$.

Let μ_n be the image on $S\times S$ of the measure $\mathbb{P}$ under the map $x\to(f_{nk}(x),x)$ so that $\mu_n\in M(\mathbb{P}_n,\mathbb{P})$ for some $\mathbb{P}_n$ concentrated on the set of points $\{x_{n1},\ldots,x_{nk}\}$. Finally,

$$W(\mathbb{P}_n,\mathbb{P})\leq\int d(x,y)d\mu_n(x,y)=\int d(f_{nk}(x),x)d\mathbb{P}(x)\leq\frac{2}{n}.$$

Since $\gamma(\mathbb{P}_n,\mathbb{P})\leq W(\mathbb{P}_n,\mathbb{P})$, this finishes the proof.

□

We are finally ready to extend Theorem 49 to separable metric spaces.

Theorem 50 *(Kantorovich-Rubinstein) If (S,d) is a separable metric space then for any two distributions $\mathbb{P},\mathbb{Q}\in\mathcal{P}_1(S)$ we have $W(\mathbb{P},\mathbb{Q})=\gamma(\mathbb{P},\mathbb{Q})$.*

Proof. By previous lemma, we can approximate $\mathbb{P}$ and $\mathbb{Q}$ by $\mathbb{P}_n$ and $\mathbb{Q}_n$ concentrated on finite (hence, compact) sets. By Theorem 49, $W(\mathbb{P}_n,\mathbb{Q}_n)=\gamma(\mathbb{P}_n,\mathbb{Q}_n)$. Finally, since both W,γ are metrics,

$$\begin{aligned} W(\mathbb{P},\mathbb{Q}) &\leq W(\mathbb{P},\mathbb{P}_n)+W(\mathbb{P}_n,\mathbb{Q}_n)+W(\mathbb{Q}_n,\mathbb{Q})\\ &= W(\mathbb{P},\mathbb{P}_n)+\gamma(\mathbb{P}_n,\mathbb{Q}_n)+W(\mathbb{Q}_n,\mathbb{Q})\\ &\leq W(\mathbb{P},\mathbb{P}_n)+W(\mathbb{Q}_n,\mathbb{Q})+\gamma(\mathbb{P}_n,\mathbb{P})+\gamma(\mathbb{Q}_n,\mathbb{Q})+\gamma(\mathbb{P},\mathbb{Q}).\end{aligned}$$

Letting $n\to\infty$ proves that $W(\mathbb{P},\mathbb{Q})\leq\gamma(\mathbb{P},\mathbb{Q})$.

□

Wasserstein's distance $W_p(\mathbb{P},\mathbb{Q})$. Given $p\geq 1$, let us define the Wasserstein distance $W_p(\mathbb{P},\mathbb{Q})$ on $\mathcal{P}_p(\mathbb{R}^n)=\{\mathbb{P}:\int|x|^p d\mathbb{P}(x)<\infty\}$ corresponding to the cost function $d(x,y)=|x-y|^p$ by

$$\begin{aligned} W_p(\mathbb{P},\mathbb{Q})^p &:= \inf\Big\{\int|x-y|^p d\mu(x,y):\mu\in M(\mathbb{P},\mathbb{Q})\Big\}\\ &= \sup\Big\{\int f d\mathbb{P}+\int g d\mathbb{Q}: f(x)+g(y)<|x-y|^p\Big\}. \qquad (20.0.2)\end{aligned}$$

Even though for $p>1$ the function $d(x,y)$ is not a metric, equality in (20.0.2) for compactly supported measures $\mathbb{P}$ and $\mathbb{Q}$ follows from the proof of Theorem 49, which does not require that d is a metric. Then one can easily extend (20.0.2) to the entire space $\mathbb{R}^n$. Moreover, W_p is a metric on $\mathcal{P}_p(\mathbb{R}^n)$ which can be shown the same way as in Lemma 41. Namely, given nearly optimal $\mu\in M(\mathbb{P},\mathbb{Q})$ and $\nu\in M(\mathbb{Q},\mathbb{T})$ we can construct $(X,Y,Z)\sim M(\mathbb{P},\mathbb{Q},\mathbb{T})$ such that $(X,Y)\sim\mu$ and $(Y,Z)\sim\nu$ and, therefore,

$$W_p(\mathbb{P},\mathbb{T})\leq(\mathbb{E}|X-Z|^p)^{\frac{1}{p}}\leq(\mathbb{E}|X-Y|^p)^{\frac{1}{p}}+(\mathbb{E}|Y-Z|^p)^{\frac{1}{p}}\leq(W_p^p(\mathbb{P},\mathbb{Q})+\varepsilon)^{\frac{1}{p}}+(W_p^p(\mathbb{Q},\mathbb{T})+\varepsilon)^{\frac{1}{p}}.$$

Let $\varepsilon\downarrow 0$.

□

Section 21

Prekopa-Leindler inequality, entropy and concentration.

In this section we will make several connections between the Kantorovich-Rubinstein theorem and other classical objects. Let us start with the following classical inequality.

Theorem 51 *(Prekopa-Leindler) Consider nonnegative integrable functions $w, u, v : \mathbb{R}^n \to [0, \infty)$ such that for some $\lambda \in (0, 1)$,*

$$w(\lambda x + (1-\lambda)y) \geq u(x)^\lambda v(y)^{1-\lambda} \quad \text{for all} \quad x, y \in \mathbb{R}^n.$$

Then,

$$\int w dx \geq \Big(\int u dx\Big)^\lambda \Big(\int v dx\Big)^{1-\lambda}.$$

Proof. The proof will proceed by induction on n. Let us first show the induction step. Suppose the statement holds for n and we would like to show it for $n+1$. By assumption, for any $x, y \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$

$$w(\lambda x + (1-\lambda)y, \lambda a + (1-\lambda)b) \geq u(x,a)^\lambda v(y,b)^{1-\lambda}.$$

Let us fix a and b and consider functions

$$w_1(x) = w(x, \lambda a + (1-\lambda)b), \;\; u_1(x) = u(x,a), \;\; v_1(x) = v(x,b)$$

on $\mathbb{R}^n$ that satisfy

$$w_1(\lambda x + (1-\lambda)y) \geq u_1(x)^\lambda v_1(y)^{1-\lambda}.$$

By induction assumption,

$$\int_{\mathbb{R}^n} w_1 dx \geq \Big(\int_{\mathbb{R}^n} u_1 dx\Big)^\lambda \Big(\int_{\mathbb{R}^n} v_1 dx\Big)^{1-\lambda}.$$

These integrals still depend on a and b and we can define

$$w_2(\lambda a + (1-\lambda)b) = \int_{\mathbb{R}^n} w_1 dx = \int_{\mathbb{R}^n} w(x, \lambda a + (1-\lambda)b) dx$$

and, similarly,

$$u_2(a) = \int_{\mathbb{R}^n} u_1(x,a) dx, \;\; v_2(b) = \int_{\mathbb{R}^n} v_1(x,b) dx$$

so that

$$w_2(\lambda a + (1-\lambda)b) \geq u_2(a)^\lambda v_2(b)^{1-\lambda}.$$

These functions are defined on $\mathbb{R}$ and, by induction assumption,

$$\int_{\mathbb{R}} w_2 ds \geq \Big(\int_{\mathbb{R}} u_2 ds\Big)^{\lambda}\Big(\int_{\mathbb{R}} v_2 ds\Big)^{1-\lambda} \Longrightarrow \int_{\mathbb{R}^{n+1}} w dz \geq \Big(\int_{\mathbb{R}^{n+1}} u dz\Big)^{\lambda}\Big(\int_{\mathbb{R}^{n+1}} v dz\Big)^{1-\lambda},$$

which finishes the proof of the induction step. It remains to prove the case $n = 1$. Let us show two different proofs.

1. One approach is based on the Brunn-Minkowski inequality on the real line which says that, if γ is the Lebesgue measure and A, B are Borel sets on $\mathbb{R}$, then

$$\gamma(\lambda A + (1-\lambda)B) \geq \lambda\gamma(A) + (1-\lambda)\gamma(B),$$

where $A+B$ is the set addition, i.e. $A+B = \{a+b : a \in A, b \in B\}$. We can also assume that $u, v, w : \mathbb{R} \to [0,1]$ because the inequality is homogeneous to scaling. We have

$$\{w \geq a\} \supseteq \lambda\{u \geq a\} + (1-\lambda)\{v \geq a\}$$

because if $u(x) \geq a$ and $v(y) \geq a$ then, by assumption,

$$w(\lambda x + (1-\lambda)y) \geq u(x)^{\lambda} v(y)^{1-\lambda} \geq a^{\lambda} a^{1-\lambda} = a.$$

The Brunn-Minkowski inequality implies that

$$\gamma(w \geq a) \geq \lambda\gamma(u \geq a) + (1-\lambda)\gamma(v \geq a).$$

Finally,

$$\begin{aligned}
\int_{\mathbb{R}} w(z)dz &= \int_{\mathbb{R}}\int_0^1 I(x \leq w(z))dxdz = \int_0^1 \gamma(w \geq x)dx \\
&\geq \lambda\int_0^1 \gamma(u \geq x)dx + (1-\lambda)\int_0^1 \gamma(v \geq x)dx \\
&= \lambda\int_{\mathbb{R}} u(z)dz + (1-\lambda)\int_{\mathbb{R}} v(z)dz \geq \Big(\int_{\mathbb{R}} u(z)dz\Big)^{\lambda}\Big(\int_{\mathbb{R}} v(z)dz\Big)^{1-\lambda}.
\end{aligned}$$

2. Another approach is based on the transportation of measure. We can assume that $\int u = \int v = 1$ by rescaling

$$u \to \frac{u}{\int u}, \;\; v \to \frac{v}{\int v}, \;\; w \to \frac{w}{(\int u)^{\lambda}(\int v)^{1-\lambda}}.$$

Then we need to show that $\int w \geq 1$. Without loss of generality, let us assume that $u, v \geq 0$ are smooth and strictly positive, since one can easily reduce to this case. Define $x(t), y(t)$ for $0 \leq t \leq 1$ by

$$\int_{-\infty}^{x(t)} u(s)ds = t, \;\; \int_{-\infty}^{y(t)} v(s)ds = t.$$

Then

$$u(x(t))x'(t) = 1, \;\; u(y(t))y'(t) = 1$$

and the derivatives $x'(t), y'(t) > 0$. Define $z(t) = \lambda x(t) + (1-\lambda)y(t)$. Then

$$\int_{-\infty}^{+\infty} w(s)ds = \int_0^1 w(z(s))dz(s) = \int_0^1 w(\lambda x(s) + (1-\lambda)y(s))z'(s)ds.$$

By arithmetic-geometric mean inequality

$$z'(s) = \lambda x'(s) + (1-\lambda)y'(s) \geq (x'(s))^{\lambda}(y'(s))^{1-\lambda}$$

and, by assumption,

$$w(\lambda x(s) + (1-\lambda)y(s)) \geq u(x(s))^{\lambda} v(y(s))^{1-\lambda}.$$

Therefore,

$$\int w(s)ds \geq \int_0^1 \Big(u(x(s))x'(s)\Big)^{\lambda} \Big(v(y(s))y'(s)\Big)^{1-\lambda} ds = \int_0^1 1ds = 1.$$

This finishes the proof of theorem.

□

Entropy and the Kullback-Leibler divergence. Consider a probability measure $\mathbb{P}$ on $\mathbb{R}^n$ and a nonnegative measurable function $u : \mathbb{R}^n \to [0, \infty)$.

Definition (Entropy) *We define the* entropy *of u with respect to $\mathbb{P}$ by*

$$\mathbf{Ent}_{\mathbb{P}}(u) = \int u \log u d\mathbb{P} - \int u d\mathbb{P} \cdot \log \int u d\mathbb{P}.$$

One can give a different representation of entropy by

$$\mathbf{Ent}_{\mathbb{P}}(u) = \sup\Big\{ \int uv d\mathbb{P} : \int e^v d\mathbb{P} \leq 1 \Big\}. \tag{21.0.1}$$

Indeed, if we consider a convex set $V = \{v : \int e^v d\mathbb{P} \leq 1\}$ then the above supremum is obviously a solution of the following saddle point problem:

$$L(v, \lambda) = \int uv d\mathbb{P} - \lambda \Big(\int e^v d\mathbb{P} - 1 \Big) \to \sup_v \inf_{\lambda \geq 0} .$$

The functional L is linear in λ and concave in v. Therefore, by the minimax theorem, a saddle point solution exists and $\sup \inf = \inf \sup$. The integral

$$\int uv d\mathbb{P} - \lambda \int e^v d\mathbb{P} = \int (uv - \lambda e^v) d\mathbb{P}$$

can be maximized pointwise by taking v such that $u = \lambda e^v$. Then

$$L(v, \lambda) = \int u \log \frac{u}{\lambda} d\mathbb{P} - \int u d\mathbb{P} + \lambda$$

and maximizing over λ gives $\lambda = \int u$ and $v = \log(u/\int u)$. This proves (21.0.1). Suppose now that a law $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ and denote its Radon-Nikodym derivative by

$$u = \frac{d\mathbb{Q}}{d\mathbb{P}}. \tag{21.0.2}$$

Definition (Kullback-Leibler divergence) *The quantity*

$$D(\mathbb{Q}||\mathbb{P}) := \int \log u \, d\mathbb{Q} = \int \log \frac{d\mathbb{Q}}{d\mathbb{P}} \, d\mathbb{Q}$$

is called the Kullback-Leibler divergence *between $\mathbb{P}$ and $\mathbb{Q}$.*

Clearly, $D(\mathbb{Q}||\mathbb{P}) = \mathbf{Ent}_{\mathbb{P}}(u)$, since

$$\mathbf{Ent}_{\mathbb{P}}(u) = \int \log \frac{d\mathbb{Q}}{d\mathbb{P}} \cdot \frac{d\mathbb{Q}}{d\mathbb{P}} \, d\mathbb{P} - \int \frac{d\mathbb{Q}}{d\mathbb{P}} \, d\mathbb{P} \cdot \log \int \frac{d\mathbb{Q}}{d\mathbb{P}} \, d\mathbb{P} = \int \log \frac{d\mathbb{Q}}{d\mathbb{P}} \, d\mathbb{Q}.$$

The variational characterization (21.0.1) implies that

$$\text{if} \quad \int e^v d\mathbb{P} \leq 1 \quad \text{then} \quad \int v d\mathbb{Q} = \int uv d\mathbb{P} \leq D(Q||P). \tag{21.0.3}$$

Transportation inequality for log-concave measures. Suppose that a probability distribution $\mathbb{P}$ on $\mathbb{R}^n$ has the Lebesgue density $e^{-V(x)}$ where $V(x)$ is strictly convex in the following sense:

$$tV(x)+(1-t)V(y)-V(tx+(1-t)y)\geq C_p(1-t+\mathbf{o}(1-t))|x-y|^p \tag{21.0.4}$$

as $t\to 1$ for some $p\geq 2$ and $C_p>0$.

Example. One example of the distribution that satisfies (21.0.4) is the non-degenerate normal distribution $N(0,C)$ that corresponds to

$$V(x)=\frac{1}{2}(C^{-1}x,x)+\text{const}$$

for some covariance matrix C, $\det C\neq 0$. If we denote $A=C^{-1}/2$ then

$$\begin{aligned} &t(Ax,x)+(1-t)(Ay,y)-(A(tx+(1-t)y),(tx+(1-t)y))\\ &\quad=t(1-t)(A(x-y),(x-y))\geq\frac{1}{2\lambda_{\max}(C)}t(1-t)|x-y|^2, \end{aligned}\tag{21.0.5}$$

where $\lambda_{\max}(C)$ is the largest eigenvalue of C. Thus, (21.0.4) holds with $p=2$ and $C_p=1/(2\lambda_{\max}(C))$.

□

Let us prove the following useful inequality for the Wasserstein distance.

Theorem 52 *If $\mathbb{P}$ satisfies (21.0.4) and $\mathbb{Q}$ is absolutely continuous w.r.t. $\mathbb{P}$ then*

$$W_p(\mathbb{Q},\mathbb{P})^p\leq\frac{1}{C_p}D(\mathbb{Q}\|\mathbb{P}).$$

Proof. Take functions $f,g\in C(\mathbb{R}^n)$ such that

$$f(x)+g(y)\leq\frac{1}{t(1-t)}C_p(1-t+\mathbf{o}(1-t))|x-y|^p.$$

Then, by (21.0.4),

$$f(x)+g(y)\leq\frac{1}{t(1-t)}\Big(tV(x)+(1-t)V(y)-V(tx+(1-t)y)\Big)$$

and

$$t(1-t)f(x)-tV(x)+t(1-t)g(y)-(1-t)V(y)\leq -V(tx+(1-t)y).$$

This implies that

$$w(tx+(1-t)y)\geq u(x)^t v(y)^{1-t}$$

for

$$u(x)=e^{(1-t)f(x)-V(x)},\ v(y)=e^{tg(y)-V(y)}\ \text{ and }\ w(z)=e^{-V(z)}.$$

By the Prekopa-Leindler inequality,

$$\Big(\int e^{(1-t)f(x)-V(x)}dx\Big)^t\Big(\int e^{tg(x)-V(x)}dx\Big)^{1-t}\leq\int e^{-V(x)}dx$$

and since e^{-V} is the density of $\mathbb{P}$ we get

$$\Big(\int e^{(1-t)f}d\mathbb{P}\Big)^t\Big(\int e^{tg}d\mathbb{P}\Big)^{1-t}\leq 1\ \text{ and }\ \Big(\int e^{(1-t)f}d\mathbb{P}\Big)^{\frac{1}{1-t}}\Big(\int e^{tg}d\mathbb{P}\Big)^{\frac{1}{t}}\leq 1.$$

It is a simple calculus exercise to show that

$$\lim_{s\to 0}\Big(\int e^{sf}d\mathbb{P}\Big)^{\frac{1}{s}}=e^{\int f d\mathbb{P}},$$

and, therefore, letting $t \to 1$ proves that

$$\text{if} \quad f(x) + g(y) \le C_p|x-y|^p \quad \text{then} \quad \int e^g d\mathbb{P} \cdot e^{\int f d\mathbb{P}} \le 1.$$

If we denote $v = g + \int f d\mathbb{P}$ then the last inequality is $\int e^v d\mathbb{P} \le 1$ and (21.0.3) implies that

$$\int v d\mathbb{Q} = \int f d\mathbb{P} + \int g d\mathbb{Q} \le D(\mathbb{Q}||\mathbb{P}).$$

Finally, using the Kantorovich-Rubinstein theorem, (20.0.2), we get

$$\begin{aligned} W_p(\mathbb{Q},\mathbb{P})^p &= \frac{1}{C_p} \inf\left\{ \int C_p|x-y|^p d\mu(x,y) : \mu \in M(\mathbb{P},\mathbb{Q}) \right\} \\ &= \frac{1}{C_p} \sup\left\{ \int f d\mathbb{P} + \int g d\mathbb{Q} : f(x)+g(y) \le C_p|x-y|^p \right\} \le \frac{1}{C_p} D(\mathbb{Q}||\mathbb{P}) \end{aligned}$$

and this finishes the proof.

□

Concentration of Gaussian measure. Applying this result to the example before Theorem 52 gives that for the non-degenerate Gaussian distribution $\mathbb{P} = N(0,C)$,

$$W_2(\mathbb{P},\mathbb{Q}) \le \sqrt{2\lambda_{\max}(C) D(\mathbb{Q}||\mathbb{P})}. \tag{21.0.6}$$

Given a measurable set $A \subseteq \mathbb{R}^n$ with $\mathbb{P}(A) > 0$, define the conditional distribution $\mathbb{P}_A$ by

$$\mathbb{P}_A(C) = \frac{\mathbb{P}(CA)}{\mathbb{P}(A)}.$$

Then, obviously, the Radon-Nikodym derivative

$$\frac{d\mathbb{P}_A}{d\mathbb{P}} = \frac{1}{\mathbb{P}(A)} \mathrm{I}_A$$

and the Kullback-Leibler divergence

$$D(\mathbb{P}_A||\mathbb{P}) = \int_A \log \frac{1}{\mathbb{P}(A)} d\mathbb{P}_A = \log \frac{1}{\mathbb{P}(A)}.$$

Since W_2 is a metric, for any two Borel sets A and B

$$W_2(\mathbb{P}_A,\mathbb{P}_B) \le W_2(\mathbb{P}_A,\mathbb{P}) + W_2(\mathbb{P}_B,\mathbb{P}) \le \sqrt{2\lambda_{\max}(C)}\Big(\sqrt{\log\frac{1}{\mathbb{P}(A)}} + \sqrt{\log\frac{1}{\mathbb{P}(B)}}\Big).$$

Suppose that the sets A and B are apart from each other by a distance t, i.e. $d(A,B) \ge t > 0$. Then any two points in the support of measures $\mathbb{P}_A$ and $\mathbb{P}_B$ are at a distance at least t from each other and the transportation distance $W_2(\mathbb{P}_A,\mathbb{P}_B) \ge t$. Therefore,

$$t \le W_2(\mathbb{P}_A,\mathbb{P}_B) \le \sqrt{2\lambda_{\max}(C)}\Big(\sqrt{\log\frac{1}{\mathbb{P}(A)}} + \sqrt{\log\frac{1}{\mathbb{P}(B)}}\Big) \le \sqrt{4\lambda_{\max}(C)\log\frac{1}{\mathbb{P}(A)\mathbb{P}(B)}}.$$

Therefore,

$$\mathbb{P}(B) \le \frac{1}{\mathbb{P}(A)} \exp\Big(-\frac{t^2}{4\lambda_{\max}(C)}\Big).$$

In particular, if $B = \{x : d(x,A) \ge t\}$ then

$$\mathbb{P}\big(d(x,A) \ge t\big) \le \frac{1}{\mathbb{P}(A)} \exp\Big(-\frac{t^2}{4\lambda_{\max}(C)}\Big).$$

If the set A is not too small, e.g. $\mathbb{P}(A) \geq 1/2$, this implies that

$$\mathbb{P}\big(d(x,A) \geq t\big) \leq 2\exp\Big(-\frac{t^2}{4\lambda_{\max}(C)}\Big).$$

This shows that the Gaussian measure is exponentially concentrated near any "large enough" set. The constant $1/4$ in the exponent is not optimal and can be replaced by $1/2$; this is just an example of application of the above ideas. The optimal result is the famous Gaussian isoperimetry,

$$\text{if } \mathbb{P}(A) = \mathbb{P}(B) \text{ for some half-space } B \text{ then } \mathbb{P}(A^t) \geq \mathbb{P}(B^t).$$

Gaussian concentration via the Prekopa-Leindler inequality. If we denote $c = 1/\lambda_{\max}(C)$ then setting $t = 1/2$ in (21.0.5),

$$V(x) + V(y) - 2V\Big(\frac{x+y}{2}\Big) \geq \frac{c}{4}|x-y|^2.$$

Given a function f on $\mathbb{R}^n$ let us define its *infimum-convolution* by

$$g(y) = \inf_x \Big(f(x) + \frac{c}{4}|x-y|^2\Big).$$

Then, for all x and y,

$$g(y) - f(x) \leq \frac{c}{4}|x-y|^2 \leq V(x) + V(y) - 2V\Big(\frac{x+y}{2}\Big). \tag{21.0.7}$$

If we define

$$u(x) = e^{-f(x)-V(x)}, \;\; v(y) = e^{g(y)-V(y)}, \;\; w(z) = e^{-V(z)}$$

then (21.0.7) implies that

$$w\Big(\frac{x+y}{2}\Big) \geq u(x)^{1/2} v(y)^{1/2}.$$

The Prekopa-Leindler inequality with $\lambda = 1/2$ implies that

$$\int e^g d\mathbb{P} \int e^{-f} d\mathbb{P} \leq 1. \tag{21.0.8}$$

Given a measurable set A, let f be equal to 0 on A and $+\infty$ on the complement of A. Then

$$g(y) = \frac{c}{4} d(x,A)^2$$

and (21.0.8) implies

$$\int \exp\Big(\frac{c}{4} d(x,A)^2\Big) d\mathbb{P}(x) \leq \frac{1}{\mathbb{P}(A)}.$$

By Chebyshev's inequality,

$$\mathbb{P}\big(d(x,A) \geq t\big) \leq \frac{1}{\mathbb{P}(A)} \exp\Big(-\frac{ct^2}{4}\Big) = \frac{1}{\mathbb{P}(A)} \exp\Big(-\frac{t^2}{4\lambda_{\max}(C)}\Big).$$

□

Trivial metric and total variation.

Definition A total variation *distance between probability measure* $\mathbb{P}$ *and* $\mathbb{Q}$ *on a measurable space* $(S, \mathcal{B})$ *is defined by*

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{B}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

Using the Hahn-Jordan decomposition, we can represent a signed measure $\mu = \mathbb{P} - \mathbb{Q}$ as $\mu = \mu^+ - \mu^-$ such that for some set $D \in \mathcal{B}$ and for any set $E \in \mathcal{B}$,

$$\mu^+(E) = \mu(ED) \geq 0 \text{ and } \mu^-(E) = -\mu(ED^c) \geq 0.$$

Therefore, for any $A \in \mathcal{B}$,

$$\mathbb{P}(A) - \mathbb{Q}(A) = \mu^+(A) - \mu^-(A) = \mu^+(AD) - \mu^-(AD^c)$$

which makes it obvious that

$$\sup_{A \in \mathcal{B}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \mu^+(D).$$

Let us describe some connections of the total variation distance to the Kullback-Leibler divergence and the Kantorovich-Rubinstein theorem. Let us start with the following simple observation.

Lemma 43 *If f is a measurable function on S such that $|f| \leq 1$ and $\int f d\mathbb{P} = 0$ then for any $\lambda \in \mathbb{R}$,*

$$\int e^{\lambda f} d\mathbb{P} \leq e^{\lambda^2/2}.$$

Proof. Since $(1+f)/2, (1-f)/2 \in [0,1]$ and

$$\lambda f = \frac{1+f}{2}\lambda + \frac{1-f}{2}(-\lambda),$$

by convexity of e^x we get

$$e^{\lambda f} \leq \frac{1+f}{2}e^{\lambda} + \frac{1-f}{2}e^{-\lambda} = \mathrm{ch}(\lambda) + f\mathrm{sh}(\lambda).$$

Therefore,

$$\int e^{\lambda f} d\mathbb{P} \leq \mathrm{ch}(\lambda) \leq e^{\lambda^2/2},$$

where the last inequality is easy to see by Taylor's expansion.

□

Let us now consider a *trivial metric* on S given by

$$d(x,y) = I(x \neq y). \tag{21.0.9}$$

Then a 1-Lipschitz function f w.r.t. d, $\|f\|_{\mathrm{L}} \leq 1$, is defined by the condition that for all $x, y \in S$,

$$|f(x) - f(y)| \leq 1. \tag{21.0.10}$$

Formally, the Kantorovich-Rubinstein theorem in this case would state that

$$\begin{aligned} W(\mathbb{P},\mathbb{Q}) &:= \inf\Big\{\int I(x \neq y) d\mu(x,y) : \mu \in M(\mathbb{P},\mathbb{Q})\Big\} \\ &= \sup\Big\{\Big|\int f d\mathbb{Q} - \int f d\mathbb{P}\Big| : \|f\|_{\mathrm{L}} \leq 1\Big\} =: \gamma(\mathbb{P},\mathbb{Q}). \end{aligned}$$

However, since any uncountable set S is not separable w.r.t. a trivial metric d, we can not apply the Kantorovich-Rubinstein theorem directly. In this case one can use the Hahn-Jordan decomposition to show that γ coincides with the total variation distance,

$$\gamma(\mathbb{P},\mathbb{Q}) = \mathrm{TV}(\mathbb{P},\mathbb{Q})$$

and it is easy to construct a measure $\mu \in M(\mathbb{P},\mathbb{Q})$ explicitly that witnesses the above equality. We leave this as an exercise. Thus, for the trivial metric d,

$$W(\mathbb{P},\mathbb{Q}) = \gamma(\mathbb{P},\mathbb{Q}) = \mathrm{TV}(\mathbb{P},\mathbb{Q}).$$

We have the following analogue of the KL divergence bound.

Theorem 53 *If $\mathbb{Q}$ is absolutely continuous w.r.t. $\mathbb{P}$ then*

$$\mathrm{TV}(\mathbb{P},\mathbb{Q}) \leq \sqrt{2D(\mathbb{Q}||\mathbb{P})}.$$

Proof. Take f such that (21.0.10) holds. If we define $g(x) = f(x) - \int f d\mathbb{P}$ then, clearly, $|g| \leq 1$ and $\int g d\mathbb{P} = 0$. The above lemma implies that for any $\lambda \in \mathbb{R}$,

$$\int e^{\lambda f - \lambda \int f d\mathbb{P} - \lambda^2/2} d\mathbb{P} \leq 1.$$

The variational characterization of entropy (21.0.3) implies that

$$\lambda \int f d\mathbb{Q} - \lambda \int f d\mathbb{P} - \lambda^2/2 \leq D(\mathbb{Q}||\mathbb{P})$$

and for $\lambda > 0$ we get

$$\int f d\mathbb{Q} - \int f d\mathbb{P} \leq \frac{\lambda}{2} + \frac{1}{\lambda} D(\mathbb{Q}||\mathbb{P}).$$

Minimizing the right hand side over $\lambda > 0$, we get

$$\int f d\mathbb{Q} - \int f d\mathbb{P} \leq \sqrt{2D(\mathbb{Q}||\mathbb{P})}.$$

Applying this to f and $-f$ yields the result.

□

Section 22

Stochastic Processes. Brownian Motion.

We have developed a general theory of convergence of laws on (separable) metric spaces and in the following two sections we will look at some specific examples of convergence on the spaces of continuous functions $(C[0,1], \|\cdot\|_\infty)$ and $(C(\mathbb{R}_+), d)$, where d is a metric metrizing uniform convergence on compacts. These examples will describe a certain central limit theorem type results on these spaces and in this section we will define the corresponding limiting Gaussian laws, namely, the Brownian motion and Brownian bridge. We will start with basic definitions and basic regularity results in the presence of continuity. Given a set T and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a *stochastic process* is a function

$$X_t(\omega) = X(t,\omega) : T \times \Omega \to \mathbb{R}$$

such that for each $t \in T$, $X_t : \Omega \to \mathbb{R}$ is a random variable, i.e. a measurable function. In other words, a stochastic process is a collection of random variables X_t indexed by a set T. A stochastic process is often defined by specifying finite dimensional (f.d.) distributions $\mathbb{P}_F = \mathcal{L}(\{X_t\}_{t\in F})$ for all finite subsets $F \subseteq T$. Kolmogorov's theorem then guarantees the existence of a probability space on which the process is defined, under the natural consistency condition

$$F_1 \subseteq F_2 \Rightarrow \mathbb{P}_{F_1} = \mathbb{P}_{F_2}\Big|_{\mathbb{R}^{F_1}}.$$

One can also think of a process as a function on Ω with values in $\mathbb{R}^T = \{f : T \to \mathbb{R}\}$, because for a fixed $\omega \in \Omega$, $X_t(\omega) \in \mathbb{R}^T$ is a (random) function of t. In Kolmogorov's theorem, given a family of consistent f.d. distributions, a process was defined on the probability space $(\mathbb{R}^T, \mathcal{B}_T)$, where $\mathcal{B}_T$ is the cylindrical σ-algebra generated by the algebra of cylinders $B \times \mathbb{R}^{T\setminus F}$ for Borel sets B in $\mathbb{R}^F$ and all finite F. When T is uncountable, some very natural sets such as

$$\Big\{\sup_{t\in T} X_t > 1\Big\} = \bigcup_{t\in T}\{X_t > 1\}$$

might be not measurable on $\mathcal{B}_T$. However, in our examples we will deal with continuous processes that possess additional regularity properties.

Definition. If (T,d) is a metric space then a process X_t is called *sample continuous* if for all $\omega \in \Omega$, $X_t(\omega) \in C(T,d)$ - the space of continuous function on (T,d). The process X_t is called *continuous in probability* if $X_t \to X_{t_0}$ in probability whenever $t \to t_0$.

Example. Let $T = [0,1]$, $(\Omega, \mathbb{P}) = ([0,1], \lambda)$ where λ is the Lebesgue measure. Let $X_t(\omega) = \mathrm{I}(t = \omega)$ and $X'_t(\omega) = 0$. F.d. distributions of these processes are the same because for any fixed $t \in [0,1]$,

$$\mathbb{P}(X_t = 0) = \mathbb{P}(X'_t = 0) = 1.$$

However, $\mathbb{P}(X_t \text{ is continuous}) = 0$ but for X'_t this probability is 1.

□

Definition. Let (T, d) be a metric space. The process X_t is *measurable* if

$$X_t(\omega) : T \times \Omega \to \mathbb{R}$$

is jointly measurable on the product space $(T, \mathcal{B}) \times (\Omega, \mathcal{F})$, where $\mathcal{B}$ is the Borel σ-algebra on T.

Lemma 44 *If (T, d) is a separable metric space and X_t is sample continuous then X_t is measurable.*

Proof. Let $(S_j)_{j\geq 1}$ be a measurable partition of T such that $\text{diam}(S_j) \leq \frac{1}{n}$. For each non-empty S_j, let us take a point $t_j \in S_j$ and define

$$X_t^n(\omega) = X_{t_j}(\omega) \quad \text{for} \quad t \in S_j.$$

$X_t^n(\omega)$ is, obviously, measurable on $T \times \Omega$ because for any Borel set A on $\mathbb{R}$,

$$\Big\{(\omega, t) \,:\, X_t^n(\omega) \in A\Big\} = \bigcup_{j\geq 1} \Big\{\omega \,:\, X_{t_j}(\omega) \in A\Big\} \times S_j.$$

$X_t(w)$ is sample continuous and, therefore, $X_t^n(\omega) \to X_t(\omega)$ for all (ω, t). Hence, X is also measurable.

□

If (T, d) is a compact metric space and X_t is a sample continuous process indexed by T then we can think of X_t as an element of the metric space of continuous functions $(C(T, d), ||\cdot||_\infty)$, rather then simply an element of $\mathbb{R}^T$. We can define measurable events on this space in two different ways. On one hand, we have the natural Borel σ-algebra $\mathcal{B}$ on $C(T)$ generated by the open (or closed) balls

$$B_g(\varepsilon) = \{f \in C(T) : ||f - g||_\infty < \varepsilon\}.$$

On the other hand, if we think of $C(T)$ as a subspace of $\mathbb{R}^T$, we can consider a σ-algebra

$$S_T = \Big\{B \cap C(T) : B \in \mathcal{B}_T\Big\}$$

which is the intersection of the cylindrical σ-algebra $\mathcal{B}_T$ with $C(T)$. It turns out that these two definitions coincide. An important implication of this is that the law of any sample continuous process X_t on (T, d) is completely determined by its finite dimensional distributions.

Lemma 45 *If (T, d) is a separable metric space then $\mathcal{B} = S_T$.*

Proof. Let us first show that $S_T \subseteq \mathcal{B}$. Any element of the cylindrical algebra that generates the cylindrical σ-algebra $\mathcal{B}_T$ is given by

$$B \times \mathbb{R}^{T\setminus F} \quad \text{for a finite } F \subseteq T \text{ and for some Borel set } \quad B \subseteq \mathbb{R}^F.$$

Then

$$\big(B \times \mathbb{R}^{T\setminus F}\big) \bigcap C(T) = \Big\{x \in C(T) \,:\, (x_t)_{t\in F} \in B\Big\} = \big\{\pi_F(x) \in B\big\}$$

where $\pi_F : C(T) \to \mathbb{R}^F$ is the finite dimensional projection such that $\pi_F(x) = (x_t)_{t\in F}$. Projection π_F is, obviously, continuous in the $||\cdot||_\infty$ norm and, therefore, measurable on the Borel σ-algebra $\mathcal{B}$ generated by open sets in the $\|\cdot\|_\infty$ norm. This implies that $\big\{\pi_F(x) \in B\big\} \in \mathcal{B}$ and, thus, $S_T \subseteq \mathcal{B}$. Let us now show that $\mathcal{B} \subseteq S_T$. Let T' be a countable dense subset of T. Then, by continuity, any closed ε-ball in $C(T)$ can be written as

$$\big\{f \in C(T) : ||f - g||_\infty \leq \varepsilon\big\} = \bigcap_{t\in T'} \big\{f \in C(T) \,:\, |f(t) - g(t)| \leq \varepsilon\big\} \in S_T$$

and this finished the proof.

□

In the remainder of the section we will define two specific sample continuous stochastic processes.

Brownian motion. Brownian motion is a sample continuous process X_t on $T = \mathbb{R}_+$ such that (a) the distribution of X_t is centered Gaussian for each $t \geq 0$; (b) $X_0 = 0$ and $\mathbb{E}X_1^2 = 1$; (c) if $t < s$ then X_t and $X_s - X_t$ are independent and $\mathcal{L}(X_s - X_t) = \mathcal{L}(X_{s-t})$. If we denote $\sigma^2(t) = \mathrm{Var}(X_t)$ then these properties imply

$$\sigma^2(nt) = n\sigma^2(t), \;\; \sigma^2\Big(\frac{t}{m}\Big) = \frac{1}{m}\sigma^2(t) \;\; \text{and} \;\; \sigma^2(qt) = q\sigma^2(t)$$

for all rational q. Since $\sigma^2(1) = 1$, $\sigma^2(q) = q$ for all rational q and, by sample continuity, $\sigma^2(t) = t$ for all $t \geq 0$. Therefore, for $s < t$,

$$\mathbb{E}X_sX_t = \mathbb{E}X_s(X_s + (X_t - X_s)) = s = \min(t, s).$$

As a result, we can give an equivalent definition.

Definition. Brownian motion is a sample continuous centered Gaussian process X_t for $t \in [0, \infty)$ with the covariance $\mathrm{Cov}(X_t, X_s) = \min(t, s)$.

Without the requirement of sample continuity, the existence of such process follows from Kolmogorov's theorem since all finite dimensional distributions are consistent by construction. However, we still need to prove that there exists a sample continuous version of the process. We start with a simple estimate.

Lemma 46 *If $f(c) = \mathcal{N}(0,1)(c, \infty)$ is the tail probability of the standard normal distribution then*

$$f(c) \leq e^{-\frac{c^2}{2}} \quad \textit{for all} \quad c > 0.$$

Proof. We have

$$f(c) = \frac{1}{\sqrt{2\pi}} \int_c^\infty e^{-\frac{x^2}{2}} dx \leq \frac{1}{\sqrt{2\pi}} \int_c^\infty \frac{x}{c} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \frac{1}{c} e^{-\frac{c^2}{2}}.$$

If $c > 1/\sqrt{2\pi}$ then $f(c) \leq \exp(-c^2/2)$. If $c \leq 1/\sqrt{2\pi}$ then a simpler estimate gives the result

$$f(c) \leq f(0) = \frac{1}{2} \leq \exp\Big(-\frac{1}{2}\Big(\frac{1}{\sqrt{2\pi}}\Big)^2\Big) \leq e^{-\frac{c^2}{2}}.$$

□

Theorem 54 *There exists a continuous version of the Brownian motion.*

Proof. It is obviously enough to define X_t on the interval $[0, 1]$. Given a process X_t that has f.d. distributions of the Brownian process but is not necessarily continuous, let us define for $n \geq 1$,

$$V_k = X_{\frac{k+1}{2^n}} - X_{\frac{k}{2^n}} \quad \text{for} \quad k = 0, \ldots, 2^n - 1.$$

The variance $\mathrm{Var}(V_k) = 1/2^n$ and, by the above lemma,

$$\mathbb{P}\Big(\max_k |V_k| \geq \frac{1}{n^2}\Big) \leq 2^n \mathbb{P}\Big(|V_1| \geq \frac{1}{n^2}\Big) \leq 2^{n+1} \exp\Big(-\frac{2^{n-1}}{n^4}\Big).$$

The right hand side is summable over $n \geq 1$ and, by the Borel-Cantelli lemma,

$$\mathbb{P}\Big(\Big\{\max_k |V_k| \geq \frac{1}{n^2}\Big\} \text{ i.o.}\Big) = 0. \tag{22.0.1}$$

Given $t \in [0, 1]$ and its dyadic decomposition $t = \sum_{j=1}^\infty \frac{t_j}{2^j}$ for $t_j \in \{0, 1\}$, let us define $t(n) = \sum_{j=1}^n \frac{t_j}{2^j}$ so that

$$X_{t(n)} - X_{t(n-1)} \in \{0\} \cup \{V_k : k = 0, \ldots, 2^n - 1\}.$$

Then, the sequence

$$X_{t(n)} = 0 + \sum_{1 \leq j \leq n} (X_{t(j)} - X_{t(j-1)})$$

converges almost surely to some limit Z_t because by (22.0.1) with probability one

$$|X_{t(n)} - X_{t(n-1)}| \le n^{-2}$$

for large enough (random) $n \ge n_0(\omega)$. By construction, $Z_t = X_t$ on the dense subset of all dyadic $t \in [0,1]$. If we can prove that Z_t is sample continuous then all f.d. distributions of Z_t and X_t will coincide, which means that Z_t is a continuous version of the Brownian motion. Take any $t, s \in [0,1]$ such that $|t-s| \le 2^{-n}$. If $t(n) = \frac{k}{2^n}$ and $s(n) = \frac{m}{2^n}$, then $|k-m| \in \{0,1\}$. As a result, $|X_{t(n)} - X_{s(n)}|$ is either equal to 0 or one of the increments $|V_k|$ and, by (22.0.1), $|X_{t(n)} - X_{s(n)}| \le n^{-2}$ for large enough n. Finally,

$$\begin{aligned} |Z_t - Z_s| &\le |Z_t - X_{t(n)}| + |X_{t(n)} - X_{s(n)}| + |X_{s(n)} - Z_s| \\ &\le \sum_{l \ge n} \frac{1}{l^2} + \frac{1}{n^2} + \sum_{l \ge n} \frac{1}{l^2} \le \frac{c}{n} \end{aligned}$$

which proves the continuity of Z_t. On the event in (22.0.1) of probability zero we set $Z_t = 0$.

□

Definition. A sample continuous centered Gaussian process B_t for $t \in [0,1]$ is called a *Brownian bridge* if

$$\mathbb{E} B_t B_s = s(1-t) \quad \text{for} \quad s < t.$$

Such process exists because if X_t is a Brownian motion then $B_t = X_t - tX_1$ is a Brownian bridge, since for $s < t$,

$$\mathbb{E} B_s B_t = \mathbb{E}(X_t - tX_1)(X_s - sX_1) = s - st - ts + st = s(1-t).$$

Notice that $B_0 = B_1 = 0$.

□

Section 23

Donsker Invariance Principle.

In this section we show how the Brownian motion W_t arises in a classical central limit theorem on the space of continuous functions on $\mathbb{R}_+$. When working with continuous processes defined on $\mathbb{R}_+$, such as the Brownian motion, the metric $\|\cdot\|_\infty$ on $C(\mathbb{R}_+)$ is too strong. A more appropriate metric d can be defined by

$$d_n(f,g) = \sup_{0\le t\le n} |f(t)-g(t)| \quad \text{and} \quad d(f,g) = \sum_{n\ge 1} \frac{1}{2^n} \frac{d_n(f,g)}{1+d_n(f,g)}.$$

It is obvious that $d(f_j, f) \to 0$ if and only if $d_n(f_j, f) \to 0$ for all $n \ge 1$, i.e. d metrizes uniform convergence on compacts. $(C(\mathbb{R}_+), d)$ is also a complete separable space, since any sequence is Cauchy in d if and only if it is Cauchy for each d_n. When proving uniform tightness of laws on $(C(\mathbb{R}_+), d)$, we will need a characterization of compacts via the Arzela-Ascoli theorem, which in this case can be formulated as follows. For a function $x \in C[0,T]$, its modulus of continuity is defined by

$$m^T(x,\delta) = \sup\Big\{ |x_a - x_b| : \ |a-b| \le \delta,\ a,b \in [0,T] \Big\}.$$

Theorem 55 *(Arzela-Ascoli) A set K is compact in $(C(\mathbb{R}_+), d)$ if and only if K is closed, uniformly bounded and equicontinuous on each interval $[0,n]$. In other words,*

$$\sup_{x\in K} |x_0| < \infty \quad \textit{and} \quad \lim_{\delta\to 0} \sup_{x\in K} m^T(x,\delta) = 0 \quad \textit{for all} \quad T > 0.$$

Here is the main result about the uniform tightness of laws on $(C(\mathbb{R}_+), d)$, which is simply a translation of the Arzela-Ascoli theorem into probabilistic language.

Theorem 56 *The sequence of laws $(\mathbb{P}_n)_{n\ge 1}$ on $(C(\mathbb{R}_+), d)$ is uniformly tight if and only if*

$$\lim_{\lambda\to+\infty} \sup_{n\ge 1} \mathbb{P}_n(|x_0| > \lambda) = 0 \tag{23.0.1}$$

and

$$\lim_{\delta\downarrow 0} \sup_{n\ge 1} \mathbb{P}_n(m^T(x,\delta) > \varepsilon) = 0 \tag{23.0.2}$$

for any $T > 0$ and any $\varepsilon > 0$.

Proof. $\Longrightarrow$. For any $\gamma > 0$, there exists a compact K such that $\mathbb{P}_n(K) > 1-\gamma$ for all $n \ge 1$. By the Arzela-Ascoli theorem, $|x_0| \le \lambda$ for some $\lambda > 0$ and for all $x \in K$ and, therefore,

$$\sup_n \mathbb{P}_n(|x_0| > \lambda) \le \sup_n \mathbb{P}_n(K^c) \le \gamma.$$

Also, by equicontinuity, for any $\varepsilon > 0$ there exists $\delta_0 > 0$ such that for $\delta < \delta_0$ and for all $x \in K$ we have $m^T(x,\delta) < \varepsilon$. Therefore,

$$\sup_n \mathbb{P}_n(m^T(x,\delta) > \varepsilon) \le \sup_n \mathbb{P}_n(K^c) \le \gamma.$$

$\Longleftarrow$. Given $\gamma > 0$, find $\lambda_T > 0$ such that

$$\sup_n \mathbb{P}_n(|x_0| > \lambda_T) \le \frac{\gamma}{2^{T+1}}.$$

For each $k \ge 1$, find $\delta_k > 0$ such that

$$\sup_n \mathbb{P}_n\Big(m^T(x,\delta_k) > \frac{1}{k}\Big) \le \frac{\gamma}{2^{T+k+1}}.$$

Define a set

$$A_T = \Big\{x : |x_0| \le \lambda_T, m^T(x,\delta_k) \le \frac{1}{k} \ \text{ for all } \ k \ge 1\Big\}.$$

Then for all $n \ge 1$,

$$\mathbb{P}_n(A_T) \ge 1 - \frac{\gamma}{2^{T+1}} - \sum_k \frac{\gamma}{2^{T+k+1}} = 1 - \frac{\gamma}{2^T}.$$

By the Arzela-Ascoli theorem, the set $A = \bigcap_{T\ge1} A_T$ is compact on $(C(\mathbb{R}_+), d)$ and for all $n \ge 1$,

$$\mathbb{P}_n(A) \ge 1 - \sum_{T\ge1} \frac{\gamma}{2^T} = 1 - \gamma.$$

This proves that the sequence $(\mathbb{P}_n)$ is uniformly tight.

□

Of course, for the uniform tightness on $(C[0,1], \|\cdot\|_\infty)$ we only need the second condition (23.0.2) for $T = 1$. Also, it will be convenient to slightly relax (23.0.2) and replace it with

$$\lim_{\delta\to0} \limsup_{n\to\infty} \mathbb{P}_n(m^T(x,\delta) > \varepsilon) = 0. \tag{23.0.3}$$

Indeed, given $\gamma > 0$, find δ_0, n_0 such that for $\delta < \delta_0$ and $n > n_0$,

$$\mathbb{P}_n(m^T(x,\delta) > \varepsilon) < \gamma.$$

For each $n \le n_0$ we can find δ_n such that for $\delta < \delta_n$,

$$\mathbb{P}_n(m^T(x,\delta) > \varepsilon) < \gamma$$

because $m^T(x,\delta) \to 0$ as $\delta \to 0$ for all $x \in C(\mathbb{R}_+)$. Therefore,

$$\text{if } \ \delta < \min(\delta_0, \delta_1, \dots, \delta_{n_0}) \ \text{ then } \ \mathbb{P}_n(m^T(x,\delta) > \varepsilon) < \gamma$$

for all $n \ge 1$.

Donsker invariance principle. Let us now give a classical example of convergence on $(C(\mathbb{R}_+), d)$ to the Brownian motion W_t. Consider a sequence $(X_i)_{i\ge1}$ of i.i.d. random variables such that $\mathbb{E}X_i = 0$ and $\sigma^2 = \mathbb{E}X_i^2 < \infty$. Let us consider a continuous partial sum process on $[0,\infty)$ defined by

$$W_t^n = \frac{1}{\sqrt{n}\sigma} \sum_{i\le\lfloor nt\rfloor} X_i + (nt - \lfloor nt\rfloor)\frac{X_{\lfloor nt\rfloor+1}}{\sqrt{n}\sigma},$$

where $\lfloor nt\rfloor$ is the integer part of nt, $\lfloor nt\rfloor \le nt < \lfloor nt\rfloor + 1$. Since the last term in W_t^n is of order $n^{-1/2}$, for simplicity of notations, we will simply write

$$W_t^n = \frac{1}{\sqrt{n}\sigma} \sum_{i\le nt} X_i$$

and treat nt as an integer. By the central limit theorem,

$$\frac{1}{\sqrt{n}\sigma} \sum_{i\le nt} X_i = \sqrt{t}\frac{1}{\sqrt{nt}\sigma} \sum_{i\le nt} X_i \to \mathcal{N}(0,t).$$

Given $t < s$, we can represent

$$W_s^n = W_t^n + \frac{1}{\sqrt{n}\sigma} \sum_{nt<i\le ns} X_i$$

and since W_t^n and $W_s^n - W_t^n$ are independent, it should be obvious that the f.d. distributions of W_t^n converge to the f.d. distributions of the Brownian motion W_t. By Lemma 45, this identifies W_t as the unique possible limit of W_t^n and, if we can show that the sequence of laws $(\mathcal{L}(W_t^n))_{n\ge 1}$ is uniformly tight on $\big(C[0,\infty), d\big)$, Lemma 36 in Section 18 will imply that $W_t^n \to W_t$ weakly. Since $W_0^n = 0$, we only need to show equicontinuity (23.0.3). Let us write the modulus of continuity as

$$m^T(W^n,\delta) = \sup_{|t-s|\le\delta, t,s\in[0,T]} \Big|\frac{1}{\sqrt{n}\sigma} \sum_{ns<i\le nt} X_i\Big| \le \max_{0\le k\le nT, 0<j\le n\delta} \Big|\frac{1}{\sqrt{n}\sigma} \sum_{k<i\le k+j} X_i\Big|.$$

If instead of maximizing over all $0 \le k \le nT$, we maximize over $k = ln\delta$ for $0 \le l \le m-1$, $m := T/\delta$, i.e. in increments of $n\delta$, then it is easy to check that the maximum will decrease by at most a factor of 3, because the second maximum over $0 < j \le n\delta$ is taken over intervals of the same size $n\delta$. As a consequence, if $m^T(W^n,\delta) > \varepsilon$ then one of the events

$$\Big\{\max_{0<j\le n\delta} \Big|\frac{1}{\sqrt{n}\sigma} \sum_{ln\delta<i\le ln\delta+j} X_i\Big| > \frac{\varepsilon}{3}\Big\}$$

must occur for some $0 \le l \le m-1$. Since the number of these events is $m = T/\delta$,

$$\mathbb{P}\Big(m^T(W^n,\delta) > \varepsilon\Big) \le m\,\mathbb{P}\Big(\max_{0<j\le n\delta} \Big|\frac{1}{\sqrt{n}\sigma} \sum_{0<i\le j} X_i\Big| > \frac{\varepsilon}{3}\Big). \tag{23.0.4}$$

Kolmogorov's inequality, Theorem 11 in Section 6, implies that if $S_n = X_1 + \cdots + X_n$ and

$$\max_{0<j\le n} \mathbb{P}(|S_n - S_j| > \alpha) \le p < 1$$

then

$$\mathbb{P}\Big(\max_{0<j\le n} |S_j| > 2\alpha\Big) \le \frac{1}{1-p}\mathbb{P}(|S_n| > \alpha).$$

If we take $\alpha = \varepsilon\sqrt{n}\sigma/6$ then, by Chebyshev's inequality,

$$\mathbb{P}\Big(\Big|\sum_{j<i\le n\delta} X_i\Big| > \frac{1}{6}\varepsilon\sqrt{n}\sigma\Big) \le \frac{6^2\delta n\sigma^2}{\varepsilon^2 n\sigma^2} = 36\delta\varepsilon^{-2}$$

and, therefore, if $36\delta\varepsilon^{-2} < 1$,

$$\mathbb{P}\Big(\max_{0<j\le n\delta}\Big|\sum_{0<i\le j} X_i\Big| > \frac{1}{3}\varepsilon\sqrt{n}\sigma\Big) \le \big(1-36\delta\varepsilon^{-2}\big)^{-1}\mathbb{P}\Big(\Big|\sum_{0<i\le n\delta} X_i\Big| > \frac{\varepsilon}{6}\sqrt{n}\sigma\Big).$$

Finally, using (23.0.4) and the central limit theorem,

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}(m^T(W_t^n,\delta) > \varepsilon) &\le m\big(1-36\delta\varepsilon^{-2}\big)^{-1} \limsup_{n\to\infty} \mathbb{P}\Big(\Big|\sum_{0<i\le n\delta} X_i\Big| > \frac{\varepsilon}{6}\sqrt{n}\sigma\Big) \\
&= m\big(1-36\delta\varepsilon^{-2}\big)^{-1} 2\,\mathcal{N}(0,1)\Big(\frac{\varepsilon}{6\sqrt{\delta}},\infty\Big) \\
&\le 2T\delta^{-1}\big(1-36\delta\varepsilon^{-2}\big)^{-1}\exp\Big(-\frac{1}{2}\frac{\varepsilon^2}{6^2\delta}\Big) \to 0
\end{aligned}$$

as $\delta \to 0$. This proves that

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \mathbb{P}(m^T(W^n,\delta) > \varepsilon) = 0,$$

for all $T > 0$ and $\varepsilon > 0$ and, thus, $W_t^n \to W_t$ weakly in $(C[0,\infty), d)$.

□

Section 24

Empirical process and Kolmogorov's chaining.

Empirical process and the Kolmogorov-Smirnov test. In this sections we show how the Brownian bridge B_t arises in another central limit theorem on the space of continuous functions on $[0,1]$. Let us start with a motivating example from statistics. Suppose that $x_1, \dots, x_n$ are i.i.d. uniform random variables on $[0,1]$. By the law of large numbers, for any $t \in [0,1]$, the *empirical* c.d.f. $n^{-1}\sum_{i=1}^n \mathrm{I}(x_i \le t)$ converges to the true c.d.f. $\mathbb{P}(x_1 \le t) = t$ almost surely and, moreover, by the CLT,

$$X_t^n = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i \le t) - t\Big) \to \mathcal{N}(0, t(1-t)).$$

The stochastic process X_t^n is called the *empirical process*. The covariance of this process,

$$\mathbb{E}X_t^n X_s^n = \mathbb{E}(\mathrm{I}(x_1 \le t) - t)(\mathrm{I}(x_1 \le s) - s) = s - ts - ts + ts = s(1-t),$$

is the same as the covariance of the Brownian bridge and, by the multivariate CLT, finite dimensional distributions of the empirical process converge to f.d. distributions of the Brownian bridge,

$$\mathcal{L}\Big((X_t^n)_{t\in F}\Big) \to \mathcal{L}\Big((B_t)_{t\in F}\Big). \tag{24.0.1}$$

However, we would like to show the convergence of X_t^n to B_t in some stronger sense that would imply weak convergence of continuous functions of the process on the space $(C[0,1], \|\cdot\|_\infty)$.

The Kolmogorov-Smirnov test in statistics provides one possible motivation. Suppose that i.i.d. $(X_i)_{i\ge 1}$ have continuous distribution with c.d.f. $F(t) = \mathbb{P}(X_1 \le t)$. Let $F_n(t) = n^{-1}\sum_{i=1}^n \mathrm{I}(X_i \le t)$ be the empirical c.d.f. It is easy to see the equality in distribution

$$\sup_{t\in\mathbb{R}} \sqrt{n}|F_n(t) - F(t)| \stackrel{d}{=} \sup_{t\in[0,1]} |X_t^n|$$

because $F(X_i)$ have uniform distribution on $[0,1]$. In order to test whether $(X_i)_{i\ge1}$ come from the distribution with c.d.f. F, the statisticians need to know the distribution of the above supremum or, as approximation, the distribution of its limit. Equation (24.0.1) suggests that

$$\mathcal{L}(\sup_t |X_t^n|) \to \mathcal{L}(\sup_t |B_t|). \tag{24.0.2}$$

Since B_t is sample continuous, its distribution is the law on the metric space $(C[0,1], \|\cdot\|_\infty)$. Even though X_t^n is not continuous, its jumps are of order $n^{-1/2}$ so it has a "close" continuous version Y_t^n. Since $\|\cdot\|_\infty$ is a continuous functional on $C[0,1]$, (24.0.2) would hold if we can prove weak convergence $\mathcal{L}(Y_t^n) \to \mathcal{L}(B_t)$ on the space $(C[0,1], \|\cdot\|_\infty)$. Lemma 36 in Section 18 shows that we only need to prove uniform tightness of $\mathcal{L}(Y_t^n)$

because, by Lemma 45, (24.0.1) already identifies the law of the Brownian motion as the unique possible limit. Thus, we need to address the question of uniform tightness of $(\mathcal{L}(X_t^n))$ on the complete separable space $(C[0,1], ||\cdot||_\infty)$ or equivalently, by the result of the previous section, the equicontinuity of X_t^n,

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\Big(m(X^n,\delta)>\varepsilon\Big)=0.$$

By Chebyshev's inequality,

$$\mathbb{P}\Big(m(X^n,\delta)>\varepsilon\Big)\le\frac{1}{\varepsilon}\mathbb{E}m(X^n,\delta)$$

and we need to learn how to control $\mathbb{E}m(X^n,\delta)$. The modulus of continuity of X^n can be written as

$$\begin{aligned} m(X^n,\delta)=\sup_{|t-s|\le\delta}|X_t^n-X_s^n| &= \sqrt{n}\sup_{|t-s|\le\delta}\Big|\frac{1}{n}\sum_{i=1}^n \mathrm{I}(s<x_i\le t)-(t-s)\Big| \\ &= \sqrt{n}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(x_i)-\mathbb{E}f\Big|, \end{aligned} \tag{24.0.3}$$

where we introduced the class of functions

$$\mathcal{F}=\Big\{f(x)=\mathrm{I}(s<x\le t):\ |t-s|<\delta\Big\}. \tag{24.0.4}$$

We will develop one approach to control the expectation of (24.0.3) for general classes of functions $\mathcal{F}$ and we will only use the specific definition (24.0.4) at the very end. This will be done in several steps.

Symmetrization. At the first step, we will replace the empirical process (24.0.3) by a symmetrized version, called Rademacher process, that will be easier to control. Let $x_1',\dots,x_n'$ be independent copies of $x_1,\dots,x_n$ and let $\varepsilon_1,\dots,\varepsilon_n$ be i.i.d. *Rademacher* random variables, such that $\mathbb{P}(\varepsilon_i=1)=\mathbb{P}(\varepsilon_i=-1)=1/2$. Let us define

$$\mathbb{P}_nf=\frac{1}{n}\sum_{i=1}^n f(x_i)\quad\text{and}\quad \mathbb{P}_n'f=\frac{1}{n}\sum_{i=1}^n f(x_i').$$

Notice that $\mathbb{E}\mathbb{P}_n'f=\mathbb{E}f$. Consider the random variables

$$Z=\sup_{f\in\mathcal{F}}\big|\mathbb{P}_nf-\mathbb{E}f\big|\quad\text{and}\quad R=\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(x_i)\Big|.$$

Then, using Jensen's inequality and then triangle inequality, we can write

$$\begin{aligned} \mathbb{E}Z &= \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\mathbb{P}_nf-\mathbb{E}f\Big| = \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\mathbb{P}_nf-\mathbb{E}\mathbb{P}_n'f\Big| \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n(f(x_i)-f(x_i'))\Big| = \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(x_i)-f(x_i'))\Big| \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(x_i)\Big|+\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(x_i')\Big| = 2\mathbb{E}R. \end{aligned}$$

Equality in the second line holds because switching $x_i\leftrightarrow x_i'$ arbitrarily does not change the expectation, so the equality holds for any fixed (ε_i) and, therefore, for any random (ε_i).

□

Hoeffding's inequality. The first step to control the supremum in R is to control the sum $\sum_{i=1}^n\varepsilon_i f(x_i)$ for a fixed function f. Consider an arbitrary sequence $a_1,\dots,a_n\in\mathbb{R}$. Then the following holds.

Theorem 57 *(Hoeffding) For $t\ge 0$,*

$$\mathbb{P}\Big(\sum_{i=1}^n\varepsilon_i a_i\ge t\Big)\le\exp\Big(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\Big).$$

Proof. Given $\lambda > 0$, by Chebyshev's inequality,

$$\mathbb{P}\Big(\sum_{i=1}^n \varepsilon_i a_i \geq t\Big) \leq e^{-\lambda t}\mathbb{E}\exp\Big(\lambda\sum_{i=1}^n \varepsilon_i a_i\Big) = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\exp\big(\lambda\varepsilon_i a_i\big).$$

Using the inequality $(e^x + e^{-x})/2 \leq e^{x^2/2}$, we get

$$\mathbb{E}\exp\big(\lambda\varepsilon_i a_i\big) = \frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2} \leq \exp\Big(\frac{\lambda^2 a_i^2}{2}\Big).$$

Hence,

$$\mathbb{P}\Big(\sum_{i=1}^n \varepsilon_i a_i \geq t\Big) \leq \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^n a_i^2\Big)$$

and minimizing over $\lambda > 0$ gives the result.

□

Covering numbers, Kolmogorov's chaining and Dudley's entropy integral. To control $\mathbb{E}R$ for general classes of functions $\mathcal{F}$, we will need to use some measures of complexity of $\mathcal{F}$. First, we will show how to control the Rademacher process R conditionally on $x_1, \ldots, x_n$.

Definition. Suppose that (F, d) is a totally bounded metric space. For any $u > 0$, a u-packing number of F with respect to d is defined by

$$D(F, u, d) = \max \operatorname{card}\Big\{F_u \subseteq F : \, d(f, g) > u \text{ for all } f, g \in F_u\Big\}$$

and a u-covering number is defined by

$$N(D, u, d) = \min \operatorname{card}\Big\{F_u \subseteq F : \forall f \in F \; \exists \; g \in F_u \text{ such that } d(f, g) \leq u\Big\}.$$

□

Both packing and covering numbers measure how many points are needed to approximate any element in the set F within distance u. It is a simple exercise to show that

$$N(F, u, d) \leq D(F, u, d) \leq N(F, u/2, d)$$

and, in this sense, packing and covering numbers are closely related. Let F be a subset of the cube $[-1, 1]^n$ equipped with a scaled Euclidean metric

$$d(f, g) = \Big(\frac{1}{n}\sum_{i=1}^n (f_i - g_i)^2\Big)^{1/2}.$$

Consider the following Rademacher process on F,

$$\mathcal{R}(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f_i.$$

Then we have the following version of the classical Kolmogorov's chaining lemma.

Theorem 58 *(Kolmogorov's chaining) For any $u > 0$,*

$$\mathbb{P}\Big(\forall f \in F, \; \mathcal{R}(f) \leq 2^{9/2}\int_0^{d(0,f)} \log^{1/2} D(F, \varepsilon, d)d\varepsilon + 2^{7/2} d(0, f)\sqrt{u}\Big) \geq 1 - e^{-u}.$$

Proof. Without loss of generality, assume that $0 \in F$. Define a sequence of subsets

$$\{0\} = F_0 \subseteq F_1 \ldots \subseteq F_j \subseteq \ldots \subseteq F$$

such that F_j satisfies

1. $\forall f, g \in F_j,\ d(f,g) > 2^{-j}$,
2. $\forall f \in F$ we can find $g \in F_j$ such that $d(f,g) \le 2^{-j}$.

F_0 obviously satisfies these properties for $j = 0$. To construct F_{j+1} given F_j:

- Start with $F_{j+1} := F_j$.
- If possible, find $f \in F$ such that $d(f,g) > 2^{-(j+1)}$ for all $g \in F_{j+1}$.
- Let $F_{j+1} := F_{j+1} \cup \{f\}$ and repeat until you cannot find such f.

Define projection $\pi_j : F \to F_j$ as follows:

$$\text{for } f \in F \text{ find } g \in F_j \text{ with } d(f,g) \le 2^{-j} \text{ and set } \pi_j(f) = g.$$

Any $f \in F$ can be decomposed into the telescopic series

$$\begin{aligned} f &= \pi_0(f) + (\pi_1(f) - \pi_0(f)) + (\pi_2(f) - \pi_1(f)) + \dots \\ &= \sum_{j=1}^{\infty} (\pi_j(f) - \pi_{j-1}(f)). \end{aligned}$$

Moreover,

$$\begin{aligned} d(\pi_{j-1}(f), \pi_j(f)) &\le d(\pi_{j-1}(f), f) + d(f, \pi_j(f)) \\ &\le 2^{-(j-1)} + 2^{-j} = 3 \cdot 2^{-j} \le 2^{-j+2}. \end{aligned}$$

As a result, the jth term in the telescopic series for any $f \in F$ belongs to a finite set of possible *links*

$$L_{j-1,j} = \Big\{ f - g : f \in F_j, g \in F_{j-1}, d(f,g) \le 2^{-j+2} \Big\}.$$

Since $\mathcal{R}(f)$ is linear,

$$\mathcal{R}(f) = \sum_{j=1}^{\infty} \mathcal{R}(\pi_j(f) - \pi_{j-1}(f)).$$

We first show how to control $\mathcal{R}$ on the set of all links. Assume that $\ell \in L_{j-1,j}$. By Hoeffding's inequality,

$$\mathbb{P}\Big(\mathcal{R}(\ell) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \ell_i \ge t \Big) \le \exp\Big(-\frac{t^2}{2n^{-1} \sum_{i=1}^{n} \ell_i^2} \Big) \le \exp\Big(-\frac{t^2}{2 \cdot 2^{-2j+4}} \Big).$$

If $|F|$ denotes the cardinality of the set F then

$$|L_{j-1,j}| \le |F_{j-1}| \cdot |F_j| \le |F_j|^2$$

and, therefore,

$$\mathbb{P}\Big(\forall \ell \in L_{j-1,j},\ \mathcal{R}(\ell) \le t \Big) \ge 1 - |F_j|^2 \exp\Big(-\frac{t^2}{2^{-2j+5}} \Big) = 1 - \frac{1}{|F_j|^2} e^{-u}$$

after making a change of variables

$$t = \Big(2^{-2j+5} (4 \log |F_j| + u) \Big)^{1/2} \le 2^{7/2} 2^{-j} \log^{1/2} |F_j| + 2^{5/2} 2^{-j} \sqrt{u}.$$

Hence,

$$\mathbb{P}\Big(\forall \ell \in L_{j-1,j},\ \ \mathcal{R}(\ell) \le 2^{7/2} 2^{-j} \log^{1/2} |F_j| + 2^{5/2} 2^{-j} \sqrt{u}\ \Big) \ge 1 - \frac{1}{|F_j|^2} e^{-u}.$$

If $F_{j-1} = F_j$ then we can define $\pi_{j-1}(f) = \pi_j(f)$ and, since in this case $L_{j-1,j} = \{0\}$, there is no need to control these links. Therefore, we can assume that $|F_{j-1}| < |F_j|$ and taking a union bound for all steps,

$$\begin{aligned}
&\mathbb{P}\Big(\forall j \geq 1\ \forall \ell \in L_{j-1,j},\ \ \mathcal{R}(\ell) \leq 2^{7/2} 2^{-j} \log^{1/2} |F_j| + 2^{5/2} 2^{-j} \sqrt{u}\ \Big) \\
&\geq 1 - \sum_{j=1}^{\infty} \frac{1}{|F_j|^2} e^{-u} \geq 1 - \sum_{j=1}^{\infty} \frac{1}{(j+1)^2} e^{-u} = 1 - (\pi^2/6 - 1) e^{-u} \geq 1 - e^{-u}.
\end{aligned}$$

Given $f \in F$, let integer k be such that $2^{-(k+1)} < d(0,f) \leq 2^{-k}$. Then in the above construction we can assume that $\pi_0(f) = \ldots = \pi_k(f) = 0$, i.e. we will project f on 0 if possible. Then with probability at least $1 - e^{-u}$,

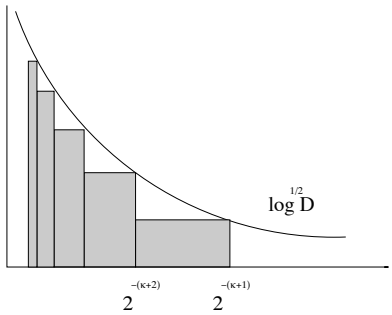
$$\begin{aligned}
\mathcal{R}(f) &= \sum_{j=k+1}^{\infty} \mathcal{R}(\pi_j(f) - \pi_{j-1}(f)) \\
&\leq \sum_{j=k+1}^{\infty} \Big(2^{7/2} 2^{-j} \log^{1/2} |F_j| + 2^{5/2} 2^{-j} \sqrt{u} \Big) \\
&\leq 2^{7/2} \sum_{j=k+1}^{\infty} 2^{-j} \log^{1/2} D(F, 2^{-j}, d) + 2^{5/2} 2^{-k} \sqrt{u}.
\end{aligned}$$

Note that $2^{-k} < 2d(f,0)$ and $2^{5/2} 2^{-k} < 2^{7/2} d(f,0)$. Finally, since packing numbers $D(F, \varepsilon, d)$ are decreasing in ε, we can write (see figure 24)

$$\begin{aligned}
2^{9/2} \sum_{j=k+1}^{\infty} 2^{-(j+1)} \log^{1/2} D(F, 2^{-j}, d) &\leq 2^{9/2} \int_0^{2^{-(k+1)}} \log^{1/2} D(F, \varepsilon, d) d\varepsilon \\
&\leq 2^{9/2} \int_0^{d(0,f)} \log^{1/2} D(F, \varepsilon, d) d\varepsilon \qquad (24.0.5)
\end{aligned}$$

since $2^{-(k+1)} < d(0,f)$. This finishes the proof.

□



The integral in (24.0.5) is called *Dudley's entropy integral.* We would like to apply the bound of the above theorem to

$$\sqrt{n} R = \sup_{f \in \mathcal{F}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(x_i) \Big|$$

for a class of functions $\mathcal{F}$ in (24.0.4). Suppose that $x_1, \ldots, x_n \in [0,1]$ are fixed and let

$$F = \Big\{ \big(f_i\big)_{1 \leq i \leq n} = \Big(\mathrm{I}\big(s < x_i \leq t\big) \Big)_{1 \leq i \leq n} :\ |t - s| \leq \delta \text{ and } t, s \in [0,1] \Big\} \subseteq \{0,1\}^n.$$

Then the following holds.

Lemma 47 *$N(F,u,d) \le Ku^{-4}$ for some absolute $K>0$ independent of the points $x_1,\dots,x_n$.*

Proof. We can assume that $x_1 \le \dots \le x_n$. Then the class F consists of all vectors of the type

$$(0\dots1\dots1\dots0),$$

i.e. the coordinates equal to 1 come in blocks. Given u, let F_u be a subset of such vectors with blocks of 1's starting and ending at the coordinates $k\lfloor nu\rfloor$. Given any vector $f\in F$, let us approximate it by a vector in $f'\in F_u$ by choosing the closest starting and ending coordinates for the blocks of 1's. The number of different coordinates will be bounded by $2\lfloor nu\rfloor$ and, therefore, the distance between f and f' will be bounded by

$$d(f,f') \le \sqrt{2n^{-1}\lfloor nu\rfloor} \le \sqrt{2u}.$$

The cardinality of F_u is, obviously, of order u^{-2}. This proves that $N(F,\sqrt{2u},d)\le Ku^{-2}$. Making the change of variables $\sqrt{2u}\to u$ proves the result.

□

To apply the Kolmogorov chaining bound to this class F let us make a simple observation that if a random variable $X\ge 0$ satisfies $\mathbb{P}(X\ge a+bt)\le Ke^{-t^2}$ for all $t\ge 0$ then

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X\ge t)dt \le a + \int_0^\infty \mathbb{P}(X\ge a+t)dt \le a + K\int_0^\infty e^{-\frac{t^2}{b^2}}dt \le a+Kb \le K(a+b).$$

Theorem 58 then implies that

$$\mathbb{E}_\varepsilon \sup_F \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f_i\Big| \le K\Big(\int_0^{D_n}\sqrt{\log\frac{K}{u}}du + D_n\Big) \tag{24.0.6}$$

where $\mathbb{E}_\varepsilon$ is the expectation with respect to (ε_i) only and

$$\begin{aligned} D_n^2 = \sup_F d(0,f)^2 &= \sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i)^2 = \sup_{|t-s|\le\delta}\frac{1}{n}\sum_{i=1}^n \mathrm{I}(s<x_i\le t) \\ &= \sup_{|t-s|\le\delta}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i\le t) - \frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i\le s)\right|. \end{aligned}$$

Since the integral on the right hand side of (24.0.6) is concave in D_n, by Jensen's inequality,

$$\mathbb{E}\sup_{\mathcal{F}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \le K\Big(\int_0^{\mathbb{E}D_n}\sqrt{\log\frac{K}{u}}du + \mathbb{E}D_n\Big).$$

By the symmetrization inequality, this finally proves that

$$\mathbb{E}m(X^n,\delta) \le K\Big(\int_0^{\mathbb{E}D_n}\sqrt{\log\frac{K}{u}}du + \mathbb{E}D_n\Big).$$

The strong law of large numbers easily implies that

$$\sup_{t\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i\le t) - t\right| \to 0 \quad \text{a.s.}$$

and, therefore, $D_n^2\to\delta$ a.s. and $\mathbb{E}D_n\to\sqrt{\delta}$. This implies that

$$\limsup_{n\to\infty}\mathbb{E}m(X_t^n,\delta) \le K\Big(\int_0^{\sqrt{\delta}}\sqrt{\log\frac{K}{u}}du + \sqrt{\delta}\Big).$$

The right-hand side goes to zero as $\delta\to 0$ and this finishes the proof of equicontinuity of X^n. As a result, for any continuous function Φ on $(C[0,1],\|\cdot\|_\infty)$ the distibution of $\Phi(X_t^n)$ converges to the distribution of $\Phi(B_t)$. For example,

$$\sqrt{n}\sup_{0\le t\le 1}\Big|\frac{1}{n}\sum \mathrm{I}(x_i\le t) - t\Big| \to \sup_{0\le t\le 1}|B_t|$$

in distribution. We will find the distribution of the right hand side in the next section.

Section 25

Markov property of Brownian motion. Reflection principles.

We showed that the empirical process converges to Brownian bridge on $(C([0,1]), \|\cdot\|_\infty)$. As a result, the distribution of a continuous function of the process will also converge, for example,

$$\sup_{0\le t\le 1} |X_t^n| \to \sup_{0\le t\le 1} |B_t|$$

weakly. We will compute the distribution of this supremum in Theorem 60 below but first we will start with a simpler example to illustrate the so called strong Markov property of the Brownian motion.

Given a process W_t on $(C[0,\infty), d)$, let $\mathcal{F}_t = \sigma(W_s; s \le t)$. A random variable τ is called a stopping time if

$$\{\tau \le t\} \in \mathcal{F}_t \quad \text{for all} \quad t \ge 0.$$

For example, a hitting time $\tau_c = \inf\{t > 0, W_t = c\}, c > 0$, is a stopping time because, by sample continuity,

$$\{\tau_c \le t\} = \bigcap_{q<c} \bigcup_{r<t} \{W_r > q\}$$

where the intersection and union are over rational numbers q, r. If W_t is the Brownian motion then strong Markov property of W_t states, informally, that the increment process $W_{\tau+t} - W_\tau$ after the stopping time is independent of the σ-algebra $\mathcal{F}_\tau$ generated by W_t up to the stopping time τ and, moreover, $W_{\tau+t} - W_\tau$ has the same distribution as W_t. This property is very similar to the property of stopping times for sums of i.i.d. random variables, in Section 7. However, to avoid subtle measure theoretic considerations, we will simply approximate arbitrary stopping times by dyadic stopping times for which Markov property can be used more directly, by summing over all possible values. If τ is a stopping time then

$$\tau_n = \frac{\lfloor 2^n \tau \rfloor + 1}{2^n}$$

is also a stopping time. Indeed, if

$$\frac{k}{2^n} \le \tau < \frac{k+1}{2^n} \quad \text{then} \quad \tau_n = \frac{k+1}{2^n}$$

and, therefore, for any $t \ge 0$, if $\frac{l}{2^n} \le t < \frac{l+1}{2^n}$ then

$$\{\tau_n \le t\} = \left\{\tau < \frac{l}{2^n}\right\} = \bigcup_{q<l/2^n} \{\tau \le q\} \in \mathcal{F}_t.$$

By construction, $\tau_n \downarrow \tau$ and, by continuity, $W_{\tau_n} \to W_\tau$ a.s. Let us demonstrate how to use Markov property for these dyadic approximations in the computation of the following probability,

$$\mathbb{P}(\sup_{t\le b} W_t \ge c) = \mathbb{P}(\tau_c \le b)$$

for $c > 0$. For dyadic approximation τ_n of τ_c, we can write

$$\begin{aligned}\mathbb{P}(\tau_n \le b, W_b - W_{\tau_n} \ge 0) &= \sum_{k\ge 0} \mathbb{P}\Big(\overbrace{\tau_n = k/2^n \le b}^{\text{in } \mathcal{F}_{k/2^n}}, \overbrace{W_b - W_{k/2^n}}^{\text{indep. of } \mathcal{F}_{k/2^n}} \ge 0\Big) \\ &= \frac{1}{2}\sum_{k\ge 0} \mathbb{P}\Big(\tau_n = \frac{k}{2^n} \le b\Big) = \frac{1}{2}\,\mathbb{P}(\tau_n \le b).\end{aligned}$$

Letting $n \to \infty$ and applying the portmanteau theorem,

$$\mathbb{P}(\tau_c \le b, W_b - W_{\tau_c} \ge 0) = \frac{1}{2}\,\mathbb{P}(\tau_c \le b),$$

since both sets $\{\tau_c = b\}$ and $\{W_b - W_{\tau_c} = 0\}$ are the sets of continuity because

$$\{\tau_c = b\} \subseteq \{W_b = c\} \quad \text{and} \quad \{W_b - W_{\tau_c} = 0\} \subseteq \{W_b = c\}$$

and $\mathbb{P}(W_b = c) = 0$. Finally, this implies that

$$\mathbb{P}(W_b \ge c) = \mathbb{P}(\tau_c \le b, W_b - W_{\tau_c} \ge 0) = \frac{1}{2}\,\mathbb{P}(\tau_c \le b) = \frac{1}{2}\,\mathbb{P}\Big(\sup_{t\le b} W_t \ge c\Big)$$

and, therefore,

$$\mathbb{P}\Big(\sup_{t\le b} W_t \ge c\Big) = \mathbb{P}(\tau_c \le b) = 2\mathcal{N}(0,b)(c,\infty) = 2\int_{c/\sqrt{b}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx. \tag{25.0.1}$$

The p.d.f. of τ_c satisfies

$$f_{\tau_c}(b) = \frac{1}{\sqrt{2\pi}} e^{-\frac{c^2}{2b}} \cdot \frac{c}{b^{3/2}} = \mathcal{O}(b^{-3/2})$$

as $b \to +\infty$, which means that $\mathbb{E}\tau_c = \infty$.

□

Reflection principles. If x_t is the Brownian motion then $y_t = x_t - tx_1$ is the Brownian bridge for $t \in [0,1]$. The next lemma shows that we can think of the Brownian bridge as the Brownian motion conditioned to be equal to zero at time $t = 1$ (pinned down Brownian motion).

Lemma 48 *Conditional distribution of x_t given $|x_1| < \varepsilon$ converges to the law of y_t,*

$$\mathcal{L}\big(x_t \big|\, |x_1| < \varepsilon\big) \to \mathcal{L}(y_t)$$

as $\varepsilon \downarrow 0$.

Proof. Notice that $y_t = x_t - tx_1$ is independent of x_1 because their covariance

$$\mathbb{E}y_t x_1 = \mathbb{E}x_t x_1 - t\mathbb{E}x_1^2 = t - t = 0.$$

Therefore, the Brownian motion can be written as a sum $x_t = y_t + tx_1$ of the Brownian bridge and independent process tx_1. Therefore, if we define a random variable η_ε with distribution $\mathcal{L}(\eta_\varepsilon) = \mathcal{L}\big(x_1 \big| |x_1| < \varepsilon\big)$ independent of y_t then

$$\mathcal{L}\big(x_t \big|\, |x_1| < \varepsilon\big) = \mathcal{L}(y_t + t\eta_\varepsilon) \to \mathcal{L}(y_t).$$

as $\varepsilon \downarrow 0$.

□

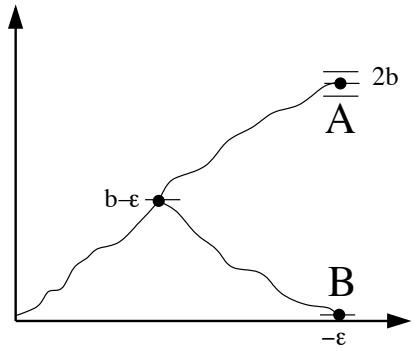


Figure 25.1: Reflecting the Brownian motion.

Theorem 59 *If y_t is the Brownian bridge then for all $b > 0$,*

$$\mathbb{P}\Big(\sup_{t\in[0,1]} y_t \geq b\Big) = e^{-2b^2}.$$

Proof. Since $y_t = x_t - tx_1$ and x_1 are independent, we can write

$$\mathbb{P}(\exists t : y_t = b) = \frac{\mathbb{P}(\exists t : x_t - tx_1 = b, |x_1| < \varepsilon)}{\mathbb{P}(|x_1| < \varepsilon)} = \frac{\mathbb{P}(\exists t : x_t = b + tx_1, |x_1| < \varepsilon)}{\mathbb{P}(|x_1| < \varepsilon)}.$$

We can estimate the numerator from below and above by

$$\mathbb{P}\big(\exists t : x_t > b + \varepsilon, |x_1| < \varepsilon\big) \leq \mathbb{P}\big(\exists t : x_t = b + tx_1, |x_1| < \varepsilon\big) \leq \mathbb{P}\big(\exists t : x_t \geq b - \varepsilon, |x_1| < \varepsilon\big).$$

Let us first analyze the upper bound. If we define a hitting time $\tau = \inf\{t : x_t = b - \varepsilon\}$ then $x_\tau = b - \varepsilon$ and

$$\mathbb{P}\big(\exists t : x_t \geq b - \varepsilon, |x_1| < \varepsilon\big) = \mathbb{P}\big(\tau \leq 1, |x_1| < \varepsilon\big) = \mathbb{P}\big(\tau \leq 1, x_1 - x_\tau \in (-b, -b + 2\varepsilon)\big).$$

For dyadic approximation as above

$$\begin{aligned}
\mathbb{P}(\tau_n \leq 1, x_1 - x_{\tau_n} \in (-b, -b + 2\varepsilon)) &= \sum_{k\geq 0} \mathbb{P}\Big(\tau_n = \frac{k}{2^n} \leq 1, x_1 - x_{k/2^n} \in (-b, -b + 2\varepsilon)\Big) \\
&= \sum_{k\geq 0} \mathbb{P}\Big(\tau_n = \frac{k}{2^n} \leq 1\Big)\mathbb{P}\Big(x_1 - x_{k/2^n} \in (-b, -b + 2\varepsilon)\Big) \\
&= \sum_{k\geq 0} \mathbb{P}\Big(\tau_n = \frac{k}{2^n} \leq 1\Big)\mathbb{P}\Big(x_1 - x_{k/2^n} \in (b - 2\varepsilon, b)\Big) \\
&= \mathbb{P}(\tau_n \leq 1, x_1 - x_{\tau_n} \in (b - 2\varepsilon, b))
\end{aligned}$$

where in the third line we used the fact that the distribution of $x_1 - x_{k/2^n}$ is symmetric around zero and, thus, we "reflected" the Brownian motion after stopping time τ as in figure 25.1. Therefore, in the limit $n \to \infty$ we get

$$\mathbb{P}\big(\exists t : x_t \geq b - \varepsilon, |x_1| < \varepsilon\big) = \mathbb{P}\big(\tau \leq 1, x_1 - x_\tau \in (b - 2\varepsilon, b)\big) = \mathbb{P}\big(x_1 \in (2b - 3\varepsilon, 2b - \varepsilon)\big)$$

because the fact that $x_1 \in (2b - 3\varepsilon, 2b - \varepsilon)$ automatically implies that $\tau \leq 1$ for $b > 0$ and ε small enough. Finally, this proves that

$$\mathbb{P}(\exists t : x_t = b) \leq \frac{\mathbb{P}(x_1 \in (2b - 3\varepsilon, 2b - \varepsilon))}{\mathbb{P}(x_1 \in (-\varepsilon, \varepsilon))} \to e^{-2b^2}$$

as $\varepsilon \to 0$. The lower bound can be analyzed similarly.

□

Theorem 60 *(Kolmogorov-Smirnov) If y_t is the Brownian bridge then for all $b > 0$,*

$$\mathbb{P}\Big(\sup_{0\le t\le 1}|y_t|\ge b\Big) = 2\sum_{n\ge 1}(-1)^{n-1}e^{-2n^2b^2}.$$

Proof. For $n \ge 1$, consider an event

$$A_n = \big\{\exists t_1 < \cdots < t_n \le 1 : y_{t_j} = (-1)^{j-1}b\big\}$$

and let τ_b and τ_{-b} be the hitting times of b and $-b$. By symmetry of the distribution of the process y_t,

$$\mathbb{P}\Big(\sup_{0\le t\le 1}|y_t|\ge b\Big) = \mathbb{P}\big(\tau_b \text{ or } \tau_{-b}\le 1\big) = 2\mathbb{P}(A_1, \tau_b < \tau_{-b}).$$

Again, by symmetry,

$$\mathbb{P}(A_n, \tau_b < \tau_{-b}) = \mathbb{P}(A_n) - \mathbb{P}(A_n, \tau_{-b} < \tau_b) = \mathbb{P}(A_n) - \mathbb{P}(A_{n+1}, \tau_b < \tau_{-b}).$$

By induction,

$$\mathbb{P}(A_1, \tau_b < \tau_{-b}) = \mathbb{P}(A_1) - \mathbb{P}(A_2) + \ldots + (-1)^{n-1}\mathbb{P}(A_n, \tau_b < \tau_{-b}).$$

As in Theorem 59, reflecting the Brownian motion each time we hit b or $-b$, one can show that

$$\mathbb{P}(A_n) = \lim_{\varepsilon\to 0}\frac{\mathbb{P}\big(x_1\in(2nb-\varepsilon, 2nb+\varepsilon)\big)}{\mathbb{P}\big(x_1\in(-\varepsilon,\varepsilon)\big)} = e^{-\frac{1}{2}(2nb)^2} = e^{-2n^2b^2}$$

and this finishes the proof. □

Given $a, b > 0$, let us compute the probability that a Brownian bridge crosses one of the levels $-a$ or b.

Theorem 61 *(Two-sided boundary) If $a, b > 0$ then*

$$\mathbb{P}(\exists t : y_t = -a \text{ or } b) = \sum_{n\ge 0}\Big(e^{-2(na+(n+1)b)^2} + e^{-2((n+1)a+nb)^2}\Big) - \sum_{n\ge 1}2e^{-2n^2(a+b)^2}. \tag{25.0.2}$$

Proof. We have

$$\mathbb{P}(\exists t : y_t = -a \text{ or } b) = \mathbb{P}(\exists t : y_t = -a, \tau_{-a} < \tau_b) + \mathbb{P}(\exists t : y_t = b, \tau_b < \tau_{-a}).$$

If we introduce the events

$$B_n = \big\{\exists t_1 < \ldots < t_n : y_{t_1} = b, y_{t_2} = -a, \ldots\big\}$$

and

$$A_n = \big\{\exists t_1 < \ldots < t_n : y_{t_1} = -a, y_{t_2} = b, \ldots\big\}$$

then, as in the previous theorem,

$$\mathbb{P}(B_n, \tau_b < \tau_{-a}) = \mathbb{P}(B_n) - \mathbb{P}(B_n, \tau_{-a} < \tau_b) = \mathbb{P}(B_n) - \mathbb{P}(A_{n+1}, \tau_{-a} < \tau_b)$$

and, similarly,

$$\mathbb{P}(A_n, \tau_{-a} < \tau_b) = \mathbb{P}(A_n) - \mathbb{P}(B_{n+1}, \tau_b < \tau_{-a}).$$

By induction,

$$\mathbb{P}(\exists t : y_t = -a \text{ or } b) = \sum_{n=1}^{\infty}(-1)^{n-1}(\mathbb{P}(A_n) + \mathbb{P}(B_n)).$$

Probabilities of the events A_n and B_n can be computed using the reflection principle as above,

$$\mathbb{P}(A_{2n}) = \mathbb{P}(B_{2n}) = e^{-2n^2(a+b)^2},\ \ \mathbb{P}(B_{2n+1}) = e^{-2(na+(n+1)b)^2},\ \ \mathbb{P}(A_{2n+1}) = e^{-2((n+1)a+nb)^2}$$

and this finishes the proof. □

If $X = -\inf y_t$ and $Y = \sup y_t$ then *the spread* of the process y_t is $\xi = X + Y$.

Theorem 62 *(Distribution of the spread) For any $t > 0$,*

$$\mathbb{P}\big(\xi \le t\big) = 1 - \sum_{n\ge 1}(8n^2t^2 - 2)e^{-2n^2t^2}.$$

Proof. First of all, (25.0.2) gives the joint c.d.f. of (X, Y) because

$$F(a,b) = \mathbb{P}(X < a, Y < b) = \mathbb{P}(-a < \inf y_t,\, \sup y_t < b) = 1 - \mathbb{P}(\exists t : y_t = -a \text{ or } b).$$

If $f(a,b) = \partial^2 F/\partial a \partial b$ is the joint p.d.f. of (X,Y) then the c.d.f of the spread $X + Y$ is

$$\mathbb{P}\big(Y + X \le t\big) = \int_0^t \int_0^{t-a} f(a,b)\, db\, da.$$

The inner integral is

$$\int_0^{t-a} f(a,b)db = \frac{\partial F}{\partial a}(a, t-a) - \frac{\partial F}{\partial a}(a,0).$$

Since

$$\begin{aligned}
\frac{\partial F}{\partial a}(a,b) &= \sum_{n\ge 0} 4n\big(na + (n+1)b\big)\, e^{-2(na+(n+1)b)^2} \\
&+ \sum_{n\ge 0} 4(n+1)\big((n+1)a + nb\big)\, e^{-2((n+1)a+nb)^2} \\
&- \sum_{n\ge 1} 8n^2(a+b)\, e^{-2n^2(a+b)^2},
\end{aligned}$$

plugging in the values $b = t - a$ and $b = 0$ gives

$$\int_0^{t-a} f(a,b)db = \sum_{n\ge 0} 4n((n+1)t - a)e^{-2((n+1)t-a)^2} + \sum_{n\ge 0} 4(n+1)(nt+a)e^{-2(nt+a)^2} - \sum_{n\ge 1} 8n^2 t e^{-2n^2t^2}.$$

Integrating over $a \in [0,t]$,

$$\begin{aligned}
\mathbb{P}\big(Y + X \le t\big) &= \sum_{n\ge 0}(2n+1)\Big(e^{-2n^2t^2} - e^{-2(n+1)^2t^2}\Big) - \sum_{n\ge 1} 8n^2t^2e^{-2n^2t^2} \\
&= 1 + 2\sum_{n\ge 1} e^{-2n^2t^2} - \sum_{n\ge 1} 8n^2t^2e^{-2n^2t^2},
\end{aligned}$$

and this finishes the proof.

□

Section 26

Laws of Brownian motion at stopping times. Skorohod's imbedding.

Let W_t be the Brownian motion.

Theorem 63 *If τ is a stopping time such that $\mathbb{E}\tau < \infty$ then $\mathbb{E}W_\tau = 0$ and $\mathbb{E}W_\tau^2 = \mathbb{E}\tau$.*

Proof. Let us start with the case when a stopping time τ takes finite number of values

$$\tau \in \{t_1, \ldots, t_n\}.$$

If $\mathcal{F}_{t_j} = \sigma\{W_t; t \le t_j\}$ then $(W_{t_j}, \mathcal{F}_{t_j})$ is a martingale since

$$\mathbb{E}(W_{t_j}|\mathcal{F}_{t_{j-1}}) = \mathbb{E}(W_{t_j} - W_{t_{j-1}} + W_{t_{j-1}}|\mathcal{F}_{t_{j-1}}) = W_{t_{j-1}}.$$

By optional stopping theorem for martingales, $\mathbb{E}W_\tau = \mathbb{E}W_{t_1} = 0$. Next, let us prove that $\mathbb{E}W_\tau^2 = \mathbb{E}\tau$ by induction on n. If $n = 1$ then $\tau = t_1$ and

$$\mathbb{E}W_\tau^2 = \mathbb{E}W_{t_1}^2 = t_1 = \mathbb{E}\tau.$$

To make an induction step from $n-1$ to n, define a stopping time $\alpha = \tau \wedge t_{n-1}$ and write

$$\mathbb{E}W_\tau^2 = \mathbb{E}(W_\alpha + W_\tau - W_\alpha)^2 = \mathbb{E}W_\alpha^2 + \mathbb{E}(W_\tau - W_\alpha)^2 + 2\mathbb{E}W_\alpha(W_\tau - W_\alpha).$$

First of all, by induction assumption, $\mathbb{E}W_\alpha^2 = \mathbb{E}\alpha$. Moreover, $\tau \neq \alpha$ only if $\tau = t_n$ in which case $\alpha = t_{n-1}$. The event

$$\{\tau = t_n\} = \{\tau \le t_{n-1}\}^c \in \mathcal{F}_{t_{n-1}}$$

and, therefore,

$$\mathbb{E}W_\alpha(W_\tau - W_\alpha) = \mathbb{E}W_{t_{n-1}}(W_{t_n} - W_{t_{n-1}})\mathrm{I}(\tau = t_n) = 0.$$

Similarly,

$$\mathbb{E}(W_\tau - W_\alpha)^2 = \mathbb{E}\mathbb{E}(\mathrm{I}(\tau = t_n)(W_{t_n} - W_{t_{n-1}})^2|\mathcal{F}_{t_{n-1}}) = (t_n - t_{n-1})\mathbb{P}(\tau = t_n).$$

Therefore,

$$\mathbb{E}W_\tau^2 = \mathbb{E}\alpha + (t_n - t_{n-1})\mathbb{P}(\tau = t_n) = \mathbb{E}\tau$$

and this finishes the proof of the induction step. Next, let us consider the case of a uniformly bounded stopping time $\tau \le M < \infty$. In the previous lecture we defined a dyadic approximation

$$\tau_n = \frac{\lfloor 2^n\tau \rfloor + 1}{2^n}$$

which is also a stopping time, $\tau_n \downarrow \tau$, and by sample continuity $W_{\tau_n} \to W_\tau$ a.s. Since (τ_n) are uniformly bounded, $\mathbb{E}\tau_n \to \mathbb{E}\tau$. To prove that $\mathbb{E}W^2_{\tau_n} \to \mathbb{E}W^2_\tau$ we need to show that the sequence $(W^2_{\tau_n})$ is uniformly integrable. Notice that $\tau_n < 2M$ and, therefore, τ_n takes possible values of the type $k/2^n$ for $k \le k_0 = \lfloor 2^n(2M) \rfloor$. Since the sequence

$$\big(W_{1/2^n}, \dots, W_{k_0/2^n}, W_{2M}\big)$$

is a martingale, adapted to a corresponding sequence of $\mathcal{F}_t$, and τ_n and $2M$ are two stopping times such that $\tau_n < 2M$, by Optional Stopping Theorem 31, $W_{\tau_n} = \mathbb{E}(W_{2M}|\mathcal{F}_{\tau_n})$. By Jensen's inequality,

$$W^4_{\tau_n} \le \mathbb{E}(W^4_{2M}|\mathcal{F}_{\tau_n}), \;\; \mathbb{E}W^4_{\tau_n} \le \mathbb{E}W^4_{2M} = 6M.$$

and uniform integrability follows by Hölder's and Chebyshev's inequalities,

$$\mathbb{E}W^2_{\tau_n}\mathrm{I}(|W_{\tau_n}| > N) \le (\mathbb{E}W^4_{\tau_n})^{1/2}(\mathbb{P}(|W_{\tau_n}| > N))^{1/2} \le \frac{6M}{N^2} \to 0$$

as $N \to \infty$, uniformly over n. This proves that $\mathbb{E}W^2_{\tau_n} \to \mathbb{E}W^2_\tau$. Since τ_n takes finite number of values, by the previous case, $\mathbb{E}W^2_{\tau_n} = \mathbb{E}\tau_n$ and letting $n \to \infty$ proves

$$\mathbb{E}W^2_\tau = \mathbb{E}\tau. \tag{26.0.1}$$

Before we consider the general case, let us notice that for two bounded stopping times $\tau \le \rho \le M$ one can similarly show that

$$\mathbb{E}(W_\rho - W_\tau)W_\tau = 0. \tag{26.0.2}$$

Namely, one can approximate the stopping times by dyadic stopping times and using that by the optional stopping theorem $(W_{\tau_n}, \mathcal{F}_{\tau_n})$, $(W_{\rho_n}, \mathcal{F}_{\rho_n})$ is a martingale,

$$\mathbb{E}(W_{\rho_n} - W_{\tau_n})W_{\tau_n} = \mathbb{E}W_{\tau_n}\big(\mathbb{E}(W_{\rho_n}|\mathcal{F}_{\tau_n}) - W_{\tau_n}\big) = 0.$$

Finally, we consider the general case. Let us define $\tau(n) = \min(\tau, n)$. For $m \le n, \tau(m) \le \tau(n)$ and

$$\mathbb{E}(W_{\tau(n)} - W_{\tau(m)})^2 = \mathbb{E}W^2_{\tau(n)} - \mathbb{E}W^2_{\tau(m)} - 2\mathbb{E}W_{\tau(m)}(W_{\tau(n)} - W_{\tau(m)}) = \mathbb{E}\tau(n) - \mathbb{E}\tau(m)$$

using (26.0.1), (26.0.2) and the fact that $\tau(n), \tau(m)$ are bounded stopping times. Since $\tau(n) \uparrow \tau$, Fatou's lemma and the monotone convergence theorem imply

$$\mathbb{E}(W_\tau - W_{\tau(m)})^2 \le \liminf_{n\to\infty}(\mathbb{E}\tau(n) - \mathbb{E}\tau(m)) = \mathbb{E}\tau - \mathbb{E}\tau(m).$$

Letting $m \to \infty$ shows that

$$\lim_{m\to\infty} \mathbb{E}(W_\tau - W_{\tau(m)})^2 = 0$$

which means that $\mathbb{E}W^2_{\tau(m)} \to \mathbb{E}W^2_\tau$. Since $\mathbb{E}W^2_{\tau(m)} = \mathbb{E}\tau(m)$ by the previous case and $\mathbb{E}\tau(m) \to \mathbb{E}\tau$ by the monotone convergence theorem, this implies that $\mathbb{E}W^2_\tau = \mathbb{E}\tau$.

□

Theorem 64 *(Skorohod's imbedding) Let Y be a random variable such that $\mathbb{E}Y = 0$ and $\mathbb{E}Y^2 < \infty$. There exists a stopping time $\tau < \infty$ such that $\mathcal{L}(W_\tau) = \mathcal{L}(Y)$.*

Proof. Let us start with the simplest case when Y takes only two values, $Y \in \{-a, b\}$ for $a, b > 0$. The condition $\mathbb{E}Y = 0$ determines the distribution of Y,

$$pb + (1-p)(-a) = 0 \quad \text{and} \quad p = \frac{a}{a+b}. \tag{26.0.3}$$

Let $\tau = \inf\{t > 0, W_t = -a \text{ or } b\}$ be a hitting time of the two-sided boundary $-a, b$. The tail probability of τ can be bounded by

$$\mathbb{P}(\tau > n) \le \mathbb{P}\big(|W_{j+1} - W_j| < a + b, 0 \le j \le n-1\big) = \mathbb{P}(|W_1| < a+b)^n = \gamma^n.$$

Therefore, $\mathbb{E}\tau < \infty$ and by the previous theorem, $\mathbb{E}W_\tau = 0$. Since $W_\tau \in \{-a, b\}$ we must have

$$\mathcal{L}(W_\tau) = \mathcal{L}(Y).$$

Let us now consider the general case. If μ is the law of Y, let us define Y by the identity $Y = Y(x) = x$ on its sample probability space $(\mathbb{R}, \mathcal{B}, \mu)$. Let us construct a sequence of σ-algebras

$$\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \ldots \subseteq \mathcal{B}$$

as follows. Let $\mathcal{B}_1$ be generated by the set $(-\infty, 0)$, i.e.

$$\mathcal{B}_1 = \big\{\emptyset, \mathbb{R}, (-\infty, 0), [0, +\infty)\big\}.$$

Given $\mathcal{B}_j$, let us define $\mathcal{B}_{j+1}$ by splitting each finite interval $[c, d) \in \mathcal{B}_j$ into two intervals $[c, (c+d)/2)$ and $[(c+d)/2, d)$ and splitting infinite interval $(-\infty, -j)$ into $(-\infty, -(j+1))$ and $[-(j+1), -j)$ and similarly splitting $[j, +\infty)$ into $[j, j+1)$ and $[j+1, \infty)$. Consider a right-closed martingale

$$Y_j = \mathbb{E}(Y|\mathcal{B}_j).$$

It is almost obvious that $\mathcal{B} = \sigma(\bigcup \mathcal{B}_j)$, which we leave as an exercise. Then, by the Levy martingale convergence, Lemma 35, $Y_j \to \mathbb{E}(Y|\mathcal{B}) = Y$ a.s. Since Y_j is measurable on $\mathcal{B}_j$, it must be constant on each simple set $[c, d) \in \mathcal{B}_j$. If $Y_j(x) = y$ for $x \in [c, d)$ then, since $Y_j = \mathbb{E}(Y|\mathcal{B}_j)$,

$$y\mu([c, d)) = \mathbb{E}Y_j \mathrm{I}_{[c,d)} = \mathbb{E}Y\mathrm{I}_{[c,d)} = \int_{[c,d)} x d\mu(x)$$

and

$$y = \frac{1}{\mu([c, d))} \int_{[c,d)} x d\mu(x). \tag{26.0.4}$$

Since in the σ-algebra $\mathcal{B}_{j+1}$ the interval $[c, d)$ is split into two intervals, the random variable Y_{j+1} can take only two values, say $y_1 < y_2$, on the interval $[c, d)$ and, since $(Y_j, \mathcal{B}_j)$ is a martingale,

$$\mathbb{E}(Y_{j+1}|\mathcal{B}_j) - Y_j = 0. \tag{26.0.5}$$

We will define stopping times τ_n such that $\mathcal{L}(W_{\tau_n}) = \mathcal{L}(Y_n)$ iteratively as follows. Since Y_1 takes only two values $-a$ and b, let $\tau_1 = \inf\{t > 0, W_t = -a \text{ or } b\}$ and we proved above that $\mathcal{L}(W_{\tau_1}) = \mathcal{L}(Y_1)$. Given τ_j define τ_{j+1} as follows:

$$\text{if } W_{\tau_j} = y \text{ for } y \text{ in (26.0.4) then } \tau_{j+1} = \inf\{t > \tau_j, W_t = y_1 \text{ or } y_2\}.$$

Let us explain why $\mathcal{L}(W_{\tau_j}) = \mathcal{L}(Y_j)$. First of all, by construction, W_{τ_j} takes the same values as Y_j. If $\mathcal{C}_j$ is the σ-algebra generated by the disjoint sets $\{W_{\tau_j} = y\}$ for y as in (26.0.4), i.e. for possible values of Y_j, then W_{τ_j} is $\mathcal{C}_j$ measurable, $\mathcal{C}_j \subseteq \mathcal{C}_{j+1}$, $\mathcal{C}_j \subseteq \mathcal{F}_{\tau_j}$ and at each step simple sets in $\mathcal{C}_j$ are split in two,

$$\{W_{\tau_j} = y\} = \{W_{\tau_{j+1}} = y_1\} \cup \{W_{\tau_{j+1}} = y_2\}.$$

By Markov's property of the Brownian motion and Theorem 63, $\mathbb{E}(W_{\tau_{j+1}} - W_{\tau_j}|\mathcal{F}_{\tau_j}) = 0$ and, therefore,

$$\mathbb{E}(W_{\tau_{j+1}}|\mathcal{C}_j) - W_{\tau_j} = 0.$$

Since on each simple set $\{W_{\tau_j} = y\}$ in $\mathcal{C}_j$, the random variable $W_{\tau_{j+1}}$ takes only two values y_1 and y_2, this equation allows us to compute the probabilities of these simple sets recursively as in (26.0.3),

$$\mathbb{P}(W_{\tau_{j+1}} = y_2) = \frac{y_2 - y}{y_2 - y_1} \mathbb{P}(W_{\tau_j} = y).$$

By (26.0.5), Y_j's satisfy the same recursive equations and this proves that $\mathcal{L}(W_{\tau_n}) = \mathcal{L}(Y_n)$. The sequence

τ_n is monotone, so it converges $\tau_n \uparrow \tau$ to some stopping time τ. Since

$$\mathbb{E}\tau_n = \mathbb{E}W_{\tau_n}^2 = \mathbb{E}Y_n^2 \leq \mathbb{E}Y^2 < \infty,$$

we have $\mathbb{E}\tau = \lim \mathbb{E}\tau_n \leq \mathbb{E}Y^2 < \infty$ and, therefore, $\tau < \infty$ a.s. Then $W_{\tau_n} \to W_\tau$ a.s. by sample continuity and since $\mathcal{L}(W_{\tau_n}) = \mathcal{L}(Y_n) \to \mathcal{L}(Y)$, this proves that $\mathcal{L}(W_\tau) = \mathcal{L}(Y)$.

□

Section 27

Laws of the Iterated Logarithm.

For convenience of notations let us denote $\ell(t) = \log\log t$.

Theorem 65 *(LIL) Let W_t be the Brownian motion and $u(t) = \sqrt{2t\ell(t)}$. Then*

$$\limsup_{t\to\infty} \frac{W_t}{u(t)} = 1.$$

Let us briefly describe the main idea that gives origin to the function $u(t)$. For $a > 1$, consider a geometric sequence $t = a^k$ and take a look at the probabilities of the following events

$$\begin{aligned} \mathbb{P}\Big(W_{a^k} \geq Lu(a^k)\Big) = \mathbb{P}\Big(\frac{W_{a^k}}{\sqrt{a^k}} \geq \frac{Lu(a^k)}{\sqrt{a^k}}\Big) &\sim \frac{1}{\sqrt{2\pi}} \frac{1}{L\sqrt{2\ell(a^k)}} \exp\Big(-\frac{1}{2}\,\frac{L^2 2a^k\ell(a^k)}{a^k}\Big) \\ &\sim \frac{1}{\sqrt{2\pi}} \frac{1}{L\sqrt{2\ell(a^k)}} \Big(\frac{1}{k\log a}\Big)^{L^2}. \end{aligned} \tag{27.0.1}$$

This series will converge or diverge depending on whether $L > 1$ or $L < 1$. Even though these events are not independent in some sense they are "almost independent" and the Borel-Cantelli lemma would imply that the upper limit of W_{a^k} behaves like $u(a^k)$. Some technical work will complete this main idea. Let us start with the following.

Lemma 49 *For any $\varepsilon > 0$,*

$$\limsup_{s\to\infty} \sup\Big\{\frac{|W_t - W_s|}{u(s)} \;:\; s \leq t \leq (1+\varepsilon)s\Big\} \leq 4\sqrt{\varepsilon} \quad a.s.$$

Proof. Let $\varepsilon, \alpha > 0$, $t_k = (1+\varepsilon)^k$ and $M_k = \alpha u(t_k)$. By symmetry, (25.0.1) and the Gaussian tail estimate in Lemma 46

$$\begin{aligned} \mathbb{P}\Big(\sup_{t_k\leq t\leq t_{k+1}} |W_t - W_{t_k}| \geq M_k\Big) &\leq 2\mathbb{P}\Big(\sup_{0\leq t\leq t_{k+1}-t_k} W_t \geq M_k\Big) \\ &= 4\mathcal{N}(0, t_{k+1}-t_k)(M_k,\infty) \leq 4\exp\Big(-\frac{1}{2}\frac{M_k^2}{(t_{k+1}-t_k)}\Big) \\ &\leq 4\exp\Big(-\frac{\alpha^2 2t_k\ell(t_k)}{2\varepsilon t_k}\Big) = 4\Big(\frac{1}{k\log(1+\varepsilon)}\Big)^{\frac{\alpha^2}{\varepsilon}}. \end{aligned}$$

If $\alpha^2 > \varepsilon$, the sum of these probabilities converges and by the Borel-Cantelli lemma, for large enough k,

$$\sup_{t_k\leq t\leq t_{k+1}} |W_t - W_{t_k}| \leq \alpha u(t_k).$$

It is easy to see that for small enough ε, $u(t_{k+1})/u(t_k) < 1+\varepsilon \le 2$. If k is such that $t_k \le s \le t_{k+1}$ then, clearly, $t_k \le s \le t \le t_{k+2}$ and, therefore, for large enough k,

$$|W_t - W_s| \le 2\alpha u(t_k) + \alpha u(t_{k+1}) \le (2\alpha + \alpha(1+\varepsilon))u(s) \le 4\alpha u(s).$$

Letting $\alpha \to \sqrt{\varepsilon}$ over some sequence finishes the proof.

□

Proof of Theorem 65. For $L = 1+\gamma > 1$, (27.0.1) and the Borel-Cantelli lemma imply that

$$W_{t_k} \le (1+\gamma)u(t_k)$$

for large enough k. If $t_k = (1+\varepsilon)^k$ then Lemma 49 implies that with probability one for large enough t (if $t_k \le t < t_{k+1}$)

$$\frac{W_t}{u(t)} = \frac{W_{t_k}}{u(t_k)}\frac{u(t_k)}{u(t)} + \frac{W_t - W_{t_k}}{u(t_k)}\frac{u(t_k)}{u(t)} \le (1+\gamma) + 4\sqrt{\varepsilon}.$$

Letting $\varepsilon, \gamma \to 0$ over some sequences proves that with probability one

$$\limsup_{t\to\infty} \frac{W_t}{u(t)} \le 1.$$

To prove that upper limit is equal to one we will use the Borel-Cantelli lemma for independent increments $W_{a^k} - W_{a^{k-1}}$ for large values of the parameter $a > 1$. If $0 < \gamma < 1$ then, similarly to (27.0.1),

$$\mathbb{P}\Big(W_{a^k} - W_{a^{k-1}} \ge (1-\gamma)u(a^k - a^{k-1})\Big) \sim \frac{1}{\sqrt{2\pi}}\frac{1}{(1-\gamma)\sqrt{2\ell(a^k - a^{k-1})}}\Big(\frac{1}{\log(a^k - a^{k-1})}\Big)^{(1-\gamma)^2}.$$

The series diverges and, since these events are independent, they occur infinitely often with probability one. We already proved (by (27.0.1)) that for $\varepsilon > 0$ for large enough k, $W_{a^k}/u(a^k) \le 1+\varepsilon$ and, therefore, by symmetry $W_{a^k}/u(a^k) \ge -(1+\varepsilon)$. This gives

$$\begin{aligned}
\frac{W_{a^k}}{u(a^k)} &\ge (1-\gamma)\frac{u(a^k - a^{k-1})}{u(a^k)} + \frac{W_{a^{k-1}}}{u(a^k)} \\
&\ge (1-\gamma)\frac{u(a^k - a^{k-1})}{u(a^k)} - (1+\varepsilon)\frac{u(a^{k-1})}{u(a^k)} \\
&= (1-\gamma)\sqrt{\frac{(a^k - a^{k-1})\ell(a^k - a^{k-1})}{a^k\ell(a^k)}} - (1+\varepsilon)\sqrt{\frac{a^{k-1}\ell(a^{k-1})}{a^k\ell(a^k)}}
\end{aligned}$$

and

$$\limsup_{t\to\infty}\frac{W_t}{u(t)} \ge \limsup_{k\to\infty}\frac{W_{a^k}}{u(a^k)} \ge (1-\gamma)\sqrt{\Big(1-\frac{1}{a}\Big)} - (1+\varepsilon)\sqrt{\frac{1}{a}}.$$

Letting $\gamma \to 0$ and $a \to \infty$ over some sequences proves that the upper limit is equal to one.

□

The LIL for Brownian motion will imply the LIL for sums of independent random variables via Skorohod's imbedding.

Theorem 66 *Suppose that $Y_1, \dots, Y_n$ are i.i.d. and $\mathbb{E}Y_i = 0, \mathbb{E}Y_i^2 = 1$. If $S_n = Y_1 + \dots + Y_n$ then*

$$\limsup_{n\to\infty}\frac{S_n}{\sqrt{2n\log\log n}} = 1 \ \textit{a.s.}$$

Proof. Let us define a stopping time $\tau(1)$ such that $W_{\tau(1)} \stackrel{\mathcal{L}}{=} Y_1$. By Markov property, the increment of the process after stopping time is independent of the process before stopping time and has the law of the Brownian motion. Therefore, we can define $\tau(2)$ such that $W_{\tau(1)+\tau(2)} - W_{\tau(1)} \stackrel{\mathcal{L}}{=} Y_2$ and, by independence,

$W_{\tau(1)+\tau(2)} \stackrel{\mathcal{L}}{=} Y_1 + Y_2$ and $\tau(1), \tau(2)$ are i.i.d. By induction, we can define i.i.d. $\tau(1), \ldots, \tau(n)$ such that $S_n \stackrel{\mathcal{L}}{=} W_{T(n)}$ where $T(n) = \tau(1) + \ldots + \tau(n)$. We have

$$\frac{S_n}{u(n)} \stackrel{\mathcal{L}}{=} \frac{W_{T(n)}}{u(n)} = \frac{W_n}{u(n)} + \frac{W_{T(n)} - W_n}{u(n)}.$$

By the LIL for the Brownian motion,

$$\limsup_{n\to\infty} \frac{W_n}{u(n)} = 1.$$

By the strong law of large numbers, $T(n)/n \to \mathbb{E}\tau(1) = \mathbb{E}Y_1^2 = 1$ a.s. For any $\varepsilon > 0$, Lemma 49 implies that for large n

$$\frac{|W_{T(n)} - W_n|}{u(n)} \le 4\sqrt{\varepsilon}$$

and letting $\varepsilon \to 0$ finishes the proof.

□

LIL for Brownian motion also implies a *local LIL*:

$$\limsup_{t\to 0} \frac{W_t}{\sqrt{2t\ell(1/t)}} = 1.$$

It is easy to check that if W_t is a Brownian motion then $tW_{1/t}$ is also the Brownian motion and the result follows by a change of variable $t \to 1/t$. To check that $tW_{1/t}$ is a Brownian motion notice that for $t < s$,

$$\mathbb{E}tW_{1/t}\Big(sW_{1/s} - tW_{1/t}\Big) = st\frac{1}{s} - t^2\frac{1}{t} = t - t = 0$$

and

$$\mathbb{E}\Big(tW_{1/t} - sW_{1/s}\Big)^2 = t + s - 2t = s - t.$$

□